A prior near-ignorance Gaussian Process model for nonparametric regression

Francesca Mangili

Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Scuola universitaria professionale della Svizzera italiana (SUPSI), Universit della Svizzera italiana (USI), Switzerland

Abstract

This paper proposes a prior near-ignorance model for regression based on a set of Gaussian Processes (GP). GPs are natural prior distributions for Bayesian regression. They offer a great modeling flexibility and have found widespread application in many regression problems. However, a GP requires the prior elicitation of its mean function, which represents our prior belief about the shape of the regression function, and of the covariance between any two function values.

In the absence of prior information, it may be difficult to fully specify these infinite dimensional parameters. In this work, by modeling the prior mean of the GP as a linear combination of a set of basis functions and assuming as prior for the combination coefficients a set of conjugate distributions obtained as limits of truncate exponential priors, we have been able to model prior ignorance about the mean of the GP. The resulting model satisfies translation invariance, learning and, under some constraints, convergence, which are desirable properties for a prior near-ignorance model. Moreover, it is shown in this paper how this model can be extended to allow for a weaker specification of the GP covariance between function values, by letting each basis function to vary in a set of functions.

Application to hypothesis testing has shown how the use of this model induces the capability of automatically detecting when a reliable decision cannot be made based on the available data.

Keywords: Gaussian Process, prior near-ignorance, nonparametric regression, hypothesis testing, Bayesian nonparametrics.

1. Introduction

A Gaussian Process (GP) extends multivariate Gaussian distributions to infinite dimensionality, thus defining a distribution over functions. Therefore, it is a natural prior distribution in Bayesian analysis for learning an unknown real-valued function $f(x)$ from a set of noisy data. GPs have found widespread use in different application domains such as classification, regression etc. [1, 2, 3, 4, 5, 6]. The reason of such success can be attributed to the great modelling flexibility of GPs, which are often used in situations where little is known about $f(x)$.

The probabilistic formulation of GPs and the simple closed form expression of their posterior inferences, makes them a good starting point to develop prior near-ignorance models for nonparametric regression. There are many techniques other than GPs available for nonparametric regression, e.g., splines, relevance vector machines, kernel smoothers, etc., some of which share strong analogies with Gaussian Processes [4, Sec. 6]. Their relative strengths and weaknesses w.r.t. GPs are discussed in [4, Sec. 7]. As they are all precise methods, we can expect them to suffer from the same weaknesses outlined below for the precise GPs.

A GP is completely specified by its mean function (encoding our prior belief about the shape of the regression function) and its kernel $k(x, x')$, used to define the covariance between any two function values: $Cov(f(x), f(x')) = k(x, x')$. A multitude of possible families exists for the covariance function, including squared exponential, polynomial, periodic, etc. (see [4]), among which the squared exponential family is by far the most popular. In the absence of prior knowledge, it can be difficult to make well grounded choices about the mean function and the kernel. A solution, proposed, among others, in [4, Ch.2.7] to allow for a weaker specification of the prior mean function, is to use a linear combination $\mathbf{h}(x)\boldsymbol{w}$ of a set of fixed basis functions $\mathbf{h}(x) = [h_1(x), \ldots, h_p(x)]$ whose coefficients $\boldsymbol{w} = [w_1, \ldots, w_p]^T$ are assumed to have an improper uniform prior distribution. Such prior belongs to the family of the so-called *non-informative* priors, which are commonly used in objective Bayesian analysis based on the fact that they satisfy some desirable invariance property, like, for instance, translation invariance. However, the improper uniform prior is just one among the priors presented in [7], all of which verify translation invariance and conjugacy with the likelihood of the GP regression model. Choosing a different prior in this family would lead to different posterior inferences. Therefore, the choice of the improper uniform prior should not be considered fully uninformative.

A way to remove this arbitrariness in the choice of the prior is to use a set of prior distributions, rather than a single distribution, and to update each of them by Bayes rule, producing a set of posterior distributions. This approach proceeds after Bayesian sensitivity analysis or Bayesian robustness [8], but with a different viewpoint, as it does not assume the existence of a correct, although unknown, prior distribution. Instead, following the theory of imprecise probabilities or coherent lower (and upper) previsions [9, 10], only upper and lower bounds for the posterior inferences of interest (expressed as expectations) are retained as valid representation of our state of knowledge. In lack of prior knowledge, to reflect this state of prior ignorance, the set of priors $\mathcal{M}$ should be as large as possible to be *vacuous* for the inferences of interest, i.e., it should provide upper and lower bounds that encompass all admissible values of such inferences. On the other side, it has to be small enough to guarantee learning from a finite number of observations. As prior ignorance and learning from data are usually conflicting properties [9, 11, 12], prior ignorance is actually required only for a limited number of basic inferences, thus modeling a state of *near*-ignorance.

In this work, we show that a regression model verifying prior near-ignorance and learning can be obtained by assuming for $\boldsymbol{w}$ the set of priors $\mathcal{M}$ presented in [7], which includes finitely additive probabilities obtained as limits of truncated exponential functions. We call this model an Imprecise GP (IGP). This set of priors can be interpreted as the set of all GPs with fixed kernel $k(x, x')$ and mean function free to vary in the set of all possible linear combinations of the set of basis functions $\boldsymbol{h}(x)$. This model improves with respect to a precise GP prior, as it models prior ignorance about the mean of the Gaussian process, i.e., about the value of the regression function. Moreover, it verifies translation invariance and, under some assumptions, convergence, which are desirable properties for a prior near-ignorance model, as discussed in [7]. Notice, however, that this model still requires to specify the covariance between any two function values, for which a good amount of prior knowledge is necessary. To address this issue, some preliminary work aimed to weaken the prior specification of the covariance is also presented. It builds on the idea of letting the basis function free to vary in a set of admissible functions, starting from the simple case of an IGP with single basis function free to vary in a set of functions obtained as linear combination of the basis functions in $\mathbf{h}(x)$.

To demonstrate the properties of the proposed approach, the IGP model has been applied to statistical hypothesis testing, focusing on a test for the difference between two regression functions given two samples of noisy observations. A nonparametric Bayesian test for the equality of regression functions based on GPs is described in [13]. In that work it is assumed that the covariates of the two samples cover the same range of values, and the comparison between the regression functions is limited to that range of values, because, having no data outside of it, nothing can be stated about the difference or equality of the two functions in other ranges of values. Instead, using the IGP it is possible to perform the equality test without worrying about the training covariate values, as the imprecise approach is able to identify those instances where the decision is prior dependent and thus it automatically detects when a reliable decision cannot be made.

2. Gaussian Process

Consider the regression model

$$y = f(x) + v, \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ and $v \sim \mathcal{N}(0, \sigma_v^2)$ is a white Gaussian noise with variance σ_v^2, and assume that we observe the data (x_i, y_i) for $i = 1, \ldots, n$. Our goal is to employ these observations to make inferences about the unknown function $f(x)$. Following the Bayesian estimation approach, we place a prior distribution on $f(x)$, and employ the observations to compute its posterior distribution; finally we use this posterior to make inferences about $f(x)$. Since $f(x)$ is a function, the Gaussian process is a natural prior distribution for it [3, 4]. Formally,

Definition 1. *Let $\mu : \mathbb{R} \to \mathbb{R}$ and $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ be a positive definite symmetric function.*[1] *A function $f(x)$ with $x \in \mathbb{R}$ is said to be distributed according to a Gaussian process with mean function $\mu(x)$ and covariance kernel $k(x, x')$ if for any finite set of covariates $x_1, \ldots, x_n$, the vector $\mathbf{f} = [f(x_1), \ldots, f(x_n)]^T$ has a multivariate n-dimensional Gaussian distribution with mean $\boldsymbol{\mu} = [\mu(x_1), \ldots, \mu(x_n)]^T$ and covariance matrix K with (i, j)-th entry equal to $k(x_i, x_j)$, $i, j = 1, \ldots, n$.*

[1] A symmetric function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is said to be positive definite if for any $\mathbf{x} = [x_1, \ldots, x_n]^T$ with $x_i \in \mathbb{R}$, the $n \times n$ matrix $[k(x_i, x_j)]_{ij}$ is positive definite.

In the following, the notation $GP(\mu(x), k_{\boldsymbol{\theta}}(x,x'))$ will denote a GP with mean function $\mu(x)$ and covariance function $k_{\boldsymbol{\theta}}(x,x')$. The subscript $\boldsymbol{\theta}$ has been introduced to highlight the fact that the covariance function usually depends on a vector of hyperparameters $\boldsymbol{\theta}$ [4]. The variables x and x' represent two generic covariates. The notation $\mathbf{x}^*$, x_i^* or x^* will be used instead hereafter to explicitly denote a finite set of test covariates at which evaluating the GP f. If $f(x) \sim GP(\mu(x), k_{\boldsymbol{\theta}}(x,x'))$, then, for any fixed m test covariates $\mathbf{x}^* = [x_1^*, \dots, x_m^*]^T$, the vector $\mathbf{f}^*$ has the Gaussian distribution

$$\mathbf{f}^*|\mathbf{x}^*, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}^*, K^{**}), \tag{2}$$

where $\boldsymbol{\mu}^* = [\mu(x_1^*), \dots, \mu(x_m^*)]^T$ and K^{**} is a $(m \times m)$-dimensional matrix with (i,j)-th entry equal to $k(x_i^*, x_j^*)$, $i,j = 1, \dots, m$. Consider a set of n inputs $\mathbf{x} = [x_1, \dots, x_n]^T$ and a vector of noisy output data $\mathbf{y} = [y_1, \dots, y_n]^T$. Based on the training data (x_i, y_i) for $i = 1, \dots, n$, and given a test input $\mathbf{x}^*$, we wish to find the posterior distribution of $\mathbf{f}^*$. From (1) and the properties of the Gaussian distribution, it follows that [4, Sec. 2.2]:

$$\left[\begin{array}{c} \mathbf{y} \\ \mathbf{f}^* \end{array} \right] \sim \mathcal{N}\left(\left[\begin{array}{c} \boldsymbol{\mu} \\ \boldsymbol{\mu}^* \end{array} \right], \left[\begin{array}{cc} K_v & K^{*T} \\ K^* & K^{**} \end{array} \right] \right), \tag{3}$$

where $K_v = K + \sigma_v^2 \mathbf{I}$, $K^* = [k_{\boldsymbol{\theta}}(x_i^*, x_j)]_{ij}$, $i = 1 \dots, m$, $j = 1 \dots, n$ and $\mathbf{I}$ is the identity matrix. When σ_v^2 is not known, it can also be considered a hyperparameter. Hence, we introduce the extended vector $\boldsymbol{\theta}_v = [\boldsymbol{\theta}, \sigma_v^2]$ of all model hyperparameters, including the noise variance. The posterior distribution of $\mathbf{f}^*$ is then

$$\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v \sim \mathcal{N}(\hat{\boldsymbol{\mu}}^*, \hat{K}^{**}), \tag{4}$$

with posterior mean vector and covariance matrix given by:

$$\hat{\boldsymbol{\mu}}^* = \boldsymbol{\mu}^* + K^* K_v^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{5}$$

$$\hat{K}^{**} = K^{**} - K^* K_v^{-1} K^{*T}. \tag{6}$$

As this hold for any finite set of test covariates $\mathbf{x}^*$, we can say that the posterior distribution of f is the GP $f(x)|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v \sim GP(\hat{\mu}(x), \hat{k}(x,x'))$ with

$$\begin{aligned} \hat{\mu}(x) &= \mu(x) + \mathbf{k}^T(x) K_v^{-1}(\mathbf{y} - \boldsymbol{\mu}), \\ \hat{k}(x,x') &= k_{\boldsymbol{\theta}}(x,x') - \mathbf{k}^T(x) K_v^{-1} \mathbf{k}(x'), \end{aligned} \tag{7}$$

where $\mathbf{k}(x) = [k_{\boldsymbol{\theta}}(x, x_1), \ldots, k_{\boldsymbol{\theta}}(x, x_n)]^T$. From this, we can compute the posterior probability $p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v)$ and make any inference about $\mathbf{f}^*$.

GP models use a kernel to define the covariance between any two function values: $Cov(f(x), f(x')) = k_{\boldsymbol{\theta}}(x, x')$. A popular choice is the squared exponential kernel:

$$k_{\boldsymbol{\theta}}(x, x') = \sigma_k^2 \exp\left[-\frac{1}{2}\frac{(x - x')^2}{\lambda^2}\right], \tag{8}$$

with hyperparameters $\boldsymbol{\theta} = (\sigma_k, \lambda) > 0$. This kernel assumes that the correlation between two function values decreases with the distance of their covariates. Observations whose covariates have a distance much larger than the lenghtscale λ are almost uncorrelated. A multitude of other possible families of covariance functions exists (polynomial, periodic, etc.), and more can be obtained by kernel composition, as positive definite kernels (i.e. those which define valid covariance functions) are closed under addition and multiplication. Once we have selected a kernel or a particular kernel composition, we must determine the values of the hyperparameters $\boldsymbol{\theta}_v$. The proper Bayesian procedure is to choose a prior for $\boldsymbol{\theta}_v$ and then determine the posterior distribution of the quantities of interest. For instance, inferences on $\mathbf{f}^*$ can be carried out by marginalizing out $\boldsymbol{\theta}_v$:

$$p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) = \textstyle\int p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v) p(\boldsymbol{\theta}_v|\mathbf{x}^*, \mathbf{x}, \mathbf{y})) d\boldsymbol{\theta}_v.$$

No closed form solution exists for $p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y})$ or for the posterior of the hyperparameters and, therefore, inferences must be computed numerically by Markov Chain Monte Carlo methods (MCMC). The convergence of MCMC methods can be quite slow when the dimension of $\boldsymbol{\theta}_v$ is high and, therefore, when we are not interested in the posterior distribution of $\boldsymbol{\theta}_v$, we can approximate the marginal of $\mathbf{f}^*$ by plugging the maximum a-posteriori (MAP) estimate for $\boldsymbol{\theta}_v$ into (4). In other words, we maximize w.r.t. $\boldsymbol{\theta}_v$ the joint marginal probability of $\mathbf{y}$ and $\boldsymbol{\theta}_v$, whose logarithm can be computed analytically [4, Ch.2]:

$$\begin{aligned}\log p(\mathbf{y}, \boldsymbol{\theta}_v|\mathbf{x}) = &-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T (K_v)^{-1} (\mathbf{y} - \boldsymbol{\mu}) + \\ &- \frac{1}{2}\log|K_v| - \frac{n}{2}\log 2\pi + \log p(\boldsymbol{\theta}_v),\end{aligned} \tag{9}$$

where $\log p(\boldsymbol{\theta}_v)$ is the prior for the vector of hyperparameters $\boldsymbol{\theta}_v$.

Often, the marginal likelihood is maximized instead of the marginal posterior probability. These maximum likelihood estimates are obtained from (9) by dropping the last term, i.e., $\log p(\boldsymbol{\theta}_v)$, and correspond to the choice of a uniform prior for the parameters in $\boldsymbol{\theta}_v$. In all the examples below we will adopt this approach.

2.1. Incorporating explicit basis functions

In practice, it can often be difficult to specify a fixed mean function and thus it is very common to use GPs with prior mean constantly equal to zero. An alternative solution, proposed, for instance, in [4, Ch.2.7], is to specify a few fixed basis functions whose coefficients $\boldsymbol{w}$ are to be inferred from the data. Consider the model

$$g(x) = f(x) + \mathbf{h}(x)\boldsymbol{w} \tag{10}$$

where $f(x) \sim GP(0, k_{\boldsymbol{\theta}}(x, x'))$ is a GP with zero mean.This formulation expresses that the data is close to a global linear model with the residuals being modeled by a GP [4]. If we take a Gaussian prior $\mathcal{N}(\mathbf{b}, B)$ on $\boldsymbol{w}$, we can integrate out the parameters $\boldsymbol{w}$ and obtain the GP

$$g(x) \sim GP(\mathbf{h}(x)\mathbf{b}, k(x, x') + \mathbf{h}(x)B\mathbf{h}^T(x')).$$

In [4, Ch 2.7] it is shown that the posterior distribution of $g(x)$ is a Gaussian process $g(x)|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v \sim GP(\hat{\mu}(x), \hat{k}(x, x'))$ with [2]

$$\begin{aligned} \hat{\mu}(x) &= \mathbf{k}(x)^T K_v^{-1}\mathbf{y} \\ &\quad + (\mathbf{h}(x)^T - H^T K_v^{-1}\mathbf{k}(x))^T (B^{-1} + H^T K_v^{-1} H)^{-1} (H^T K_v^{-1}\mathbf{y} + B^{-1}\mathbf{b}), \\ \hat{k}(x, x') &= k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}(x)^T K_v^{-1}\mathbf{k}(x') \\ &\quad + (\mathbf{h}(x)^T - H^T K_v^{-1}\mathbf{k}(x))^T (B^{-1} + H^T K_v^{-1} H)^{-1} (\mathbf{h}(x')^T - H^T K_v^{-1}\mathbf{k}(x')), \end{aligned} \tag{11}$$

where H is a $(n \times p)$-dimensional matrix with columns $[h_i(x_1), \dots, h_i(x_n)]^T$, $i = 1, \dots, p$, given by the basis functions evaluated at the training covariates. Then, given a vector $\mathbf{x}^*$ of m test covariates, the vector $\mathbf{g}^* = [g(x_1^*), \dots, g(x_m^*)]^T$ has the Gaussian distribution

$$\mathbf{g}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}^*, \hat{K}^{**}\right),$$

[2] An explicit proof of this results is not given in [4, Ch 2.7] but it can be derived from reasoning similar to the one used in the explicit feature space formulation of [4, Ch 2.1.2] or in the proof of Proposition 2 of this paper.

where $\hat{\boldsymbol{\mu}}^* = [\hat{\mu}(x_1^*), \dots, \hat{\mu}(x_m^*)]^T$ and $\hat{K}^{**}$ is a $(m \times m)$-matrix with ij entry equal to $\hat{k}(x_i^*, x_j^*)$.

In the lack of prior information it is also difficult to set a prior for $\boldsymbol{w} \in \mathcal{W}$. A solution is to select a so-called non-informative prior that satisfies, besides conjugacy, which is required for tractability of inferences, some invariance properties. It is common, for instance, to impose translation invariance.

Definition 2. *Consider the real valued bounded function $\gamma(w)$ and the group of transformations $\mathcal{F} = \{f_a(w) = w + a : a \in \mathbb{R}\}$, i.e., a shift of the parameter. Prior translation invariance is verified for γ whenever $E[\gamma(f_a)] = E[\gamma]$ for any γ and $a \in \mathbb{R}$.*

For a bounded space $\mathcal{W}$ the only prior that satisfies translation invariance is the uniform distribution. Instead, for an unbounded space, e.g., $\mathcal{W} = \mathbb{R}$ there is no countably additive probability measure that is translation invariant [7]. However, a translation invariant improper prior can be defined as the limit of uniform distributions on the interval, e.g., $\lim_{r\to\infty} \frac{1}{2r} I_{[-r,r]}$.

Using the improper uniform prior for parameter $\boldsymbol{w}$ in the GP with explicit basis functions (as suggested in [4, Ch 2.7] in case of lack of prior information) one obtains the posterior $g(x)|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v \sim GP(\hat{\mu}(x), \hat{k}(x, x'))$ with

$$\begin{aligned} \hat{\mu}(x) &= \mathbf{k}(x)^T K_v^{-1} \mathbf{y} \\ &\quad + (\mathbf{h}(x)^T - H^T K_v^{-1} \mathbf{k}(x))^T (H^T K_v^{-1} H)^{-1} (H^T K_v^{-1} \mathbf{y}), \\ \hat{k}(x, x') &= k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}(x)^T K_v^{-1} \mathbf{k}(x') \\ &\quad + (\mathbf{h}(x)^T - H^T K_v^{-1} \mathbf{k}(x))^T (H^T K_v^{-1} H)^{-1} (\mathbf{h}(x')^T - H^T K_v^{-1} \mathbf{k}(x')) \end{aligned} \tag{12}$$

The improper prior can be also interpreted as the limit of the prior $\boldsymbol{w} \sim N(\mathbf{b}, B)$ when $\mathbf{b}$ is bounded and $B^{-1} \to O_{pp}$, where O_{pp} is a $(p \times p)$-dimensional the matrix of zeros. In this view, equation (12) follows straightforwardly from (11).

2.1.1. A set of conjugate improper priors

The improper uniform prior is not the only (improper) conjugate prior that verifies translation invariance. In [7] it is shown that lower and upper expectation models defined as the limits of truncated exponential priors, that is, $\underline{E}[\gamma(w)] = \lim \inf\limits_{r\to\infty} \int \gamma(w)p(w)dw$ and $\overline{E}[\gamma(w)] = \lim \sup\limits_{r\to\infty} \int \gamma(w)p(w)dw$

where

$$p(w) = \begin{cases} \frac{\ell}{\exp(\ell r)} \exp(\ell w) I_{[-\infty,r]} & \text{if } \ell > 0, \\ \frac{-\ell}{\exp(-\ell r)} \exp(\ell w) I_{[-r,\infty]} & \text{if } \ell < 0, \\ \frac{1}{2r} I_{[-r,r]} & \text{if } \ell = 0, \end{cases}$$

and I_A is the indicator function of set A, i.e., $I_A(x) = 1$ if $x \in A$ and zero otherwise, verify translation invariance and conjugacy with any likelihood in the exponential family (with natural parameter w) and translation invariance for any sufficiently smooth function $\gamma(w)$. Notice that the prior with $\ell = 0$ corresponds to the improper uniform prior. Hereafter, for notational convenience, even in the case $\ell = 0$ these priors will be denoted as $p(w) = \frac{|\ell|}{\exp(|\ell| r)} \exp(\ell w) I_{\mathcal{W}_r}$, where

$$\mathcal{W}_r = \begin{cases} [-\infty, r] & \text{if } \ell > 0, \\ [-r, \infty] & \text{if } \ell < 0, \\ [-r, r] & \text{if } \ell = 0 \end{cases}$$

Then, we can define the limit exponential prior for the vector of coefficients $\boldsymbol{w}$ by assuming $p(w_i) = \frac{|\ell_i|}{\exp(|\ell_i|^T r_i)} \exp(\ell_i w_i) I_{\mathcal{W}_{r_i}}$. It follows that

$$p(\boldsymbol{w}) = p(w_1)p(w_2)\cdots p(w_p) = \frac{\prod_{i=1}^{p} |\ell_i|}{\exp\left(|\boldsymbol{\ell}|^T \boldsymbol{r}\right)} \exp(\boldsymbol{\ell}^T \boldsymbol{w}) \prod_{i=1}^{p} I_{\mathcal{W}_{r_i}}, \tag{13}$$

where $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_p]^T$, $\mathbf{r} = [r_1, \ldots, r_p]^T$ and $|V|$ indicates the absolute value of a matrix (or vector) V, obtained by taking the absolute value of all its elements. In order to provide posterior inferences, we need the following result to derive MAP estimates for the hyperparameters $\boldsymbol{\theta}_v$ of the GP given the limit exponential prior.

Proposition 1. *MAP estimates of the hyperparameters $\boldsymbol{\theta}_v$ given the model in (10) with base kernel $k_{\boldsymbol{\theta}}(x, x')$ and the prior $p(\boldsymbol{w}) \propto \exp\left(\boldsymbol{\ell}^T \boldsymbol{w}\right)$ are obtained by maximizing $L(\boldsymbol{\theta}_v) + \log p(\boldsymbol{\theta}_v)$ with*

$$\begin{aligned} L(\boldsymbol{\theta}_v) = &-\frac{1}{2}\mathbf{y}^T K_v^{-1}\mathbf{y} - \frac{1}{2}(H^T K_v^{-1}\mathbf{y} + \boldsymbol{\ell})^T (H^T K_v^{-1} H)^{-1} (H^T K_v^{-1}\mathbf{y} + \boldsymbol{\ell}) \\ &- \frac{1}{2}\log|K_v| - \frac{1}{2}\log|H^T K_v^{-1} H| \end{aligned} \tag{14}$$

Proof. The likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_v, \boldsymbol{w})$ for the model in (10) is

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_v, \boldsymbol{w}) = \frac{1}{(2\pi)^{\frac{n}{2}}|K_v|} \exp\left[-\frac{1}{2}(\mathbf{y} - H\boldsymbol{w})^T K_v^{-1}(\mathbf{y} - H\boldsymbol{w})\right]. \tag{15}$$

Then, the marginal likelihood given the prior in (13) is

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_v) =& \int_{\mathcal{W}} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_v, \boldsymbol{w}) d\boldsymbol{w} \\ =& T_1 \int_{\mathcal{W}_{r_1}} \dots \int_{\mathcal{W}_{r_p}} \exp\left[-\frac{1}{2}(\mathbf{y} - H\boldsymbol{w})^T K_v^{-1}(\mathbf{y} - H\boldsymbol{w}) + \boldsymbol{\ell}^T \boldsymbol{w}\right] dw_1 \dots dw_p \\ =& T_1 T_2 \int_{\mathcal{W}_{r_1}} \dots \int_{\mathcal{W}_{r_p}} \exp\left(-\frac{1}{2}(\boldsymbol{w} - \overline{\boldsymbol{w}})^T D(\boldsymbol{w} - \overline{\boldsymbol{w}})\right) dw_1 \dots dw_p, \end{aligned}$$

with $\overline{\boldsymbol{w}} = D^{-1}(H^T K_v^{-1}\mathbf{y} + \boldsymbol{\ell})$, $D = H^T K_v^{-1} H$, $T_1 = \dfrac{1}{(2\pi)^{\frac{n}{2}}|K_v|^{\frac{1}{2}}} \dfrac{\prod_{i=1}^{p} |\ell_i|}{\exp\left(|\boldsymbol{\ell}|^T \boldsymbol{r}\right)}$, and $T_2 = \exp\left(-\frac{1}{2}\mathbf{y}^T K_v^{-1}\mathbf{y} - \frac{1}{2}\overline{\boldsymbol{w}}^T D \overline{\boldsymbol{w}}\right)$. Then for $r_i \to \infty$, $i = 1, \dots, p$ the log-likelihood becomes

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_v) =& -\frac{1}{2}\left[(n-p)\log(2\pi) + \log|K_v| + \log|D| + \mathbf{y}^T K_v^{-1}\mathbf{y} \right. \\ & \left. + \overline{\boldsymbol{w}}^T D^{-1} \overline{\boldsymbol{w}}\right] + \sum_{i=1}^{p} \log|\ell_i| - \sum_{i=1}^{p} \lim_{r_i \to \infty} |\ell_i| r_i. \end{aligned}$$

As the terms $\log|\ell_i|$ and $\lim\limits_{r_i \to \infty} |\ell_i| r_i$ for all $i = 1, \dots, p$ and $(n-p)\log(2\pi)$ do not depend on the hyperparameters $\boldsymbol{\theta}_v$, maximizing $\log p(\mathbf{y}, \boldsymbol{\theta}_v) = \log p(\mathbf{y}|\boldsymbol{\theta}_v) + \log p(\boldsymbol{\theta}_v)$ is equivalent to maximize $L(\boldsymbol{\theta}_v) + \log p(\boldsymbol{\theta}_v)$, with $L(\boldsymbol{\theta}_v)$ given in (14). □

From the proof of Proposition 1 it can be noticed that if the limit exponential prior is used, then the marginal likelihood is not bounded. In general, when assuming an improper prior, the marginal likelihood cannot be used to select the covariance model by choosing, among a set of candidate kernel functions, the one that maximizes it, as often suggested in the literature about GPs.

Proposition 2. *Given the model in (10), the set of hyperparameters $\boldsymbol{\theta}_v$ and the prior $p(\boldsymbol{w}) \propto \exp(\boldsymbol{\ell}^T \boldsymbol{w})$, under the conditions $n \geq p$ and rank$(H) = p$, the* *posterior distribution of $g(x)$ is the GP $g(x)|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v \sim GP(\hat{\mu}_g(x), \hat{k}_g(x, x'))$ with*

$$\begin{aligned}\hat{\mu}_g(x) &= \mathbf{k}(x)^T K_v^{-1}\mathbf{y} \\ &\quad + (\mathbf{h}(x)^T - H^T K_v^{-1}\mathbf{k}(x))^T (H^T K_v^{-1} H)^{-1} (H^T K_v^{-1}\mathbf{y} + \boldsymbol{\ell}), \\ \hat{k}_g(x, x') &= k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}(x)^T K_v^{-1}\mathbf{k}(x') \\ &\quad + (\mathbf{h}(x)^T - H^T K_v^{-1}\mathbf{k}(x))^T (H^T K_v^{-1} H)^{-1} (\mathbf{h}(x')^T - H^T K_v^{-1}\mathbf{k}(x')).\end{aligned} \tag{16}$$

Proof. Conditioned on a fixed value of $\boldsymbol{w}$, the prior for $g(x)$ is the GP $g(x)|\boldsymbol{w} \sim GP(\mathbf{h}(x)\boldsymbol{w}, k_{\boldsymbol{\theta}}(x, x'))$. Let $\mathbf{g}^*$ be the vector of values of the function $g(x)$ evaluated at any m test covariates $\mathbf{x}^*$. Then, from (5) and (6) the posterior for $\mathbf{g}^*$ is the Gaussian distribution

$$\mathbf{g}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v, \boldsymbol{w} \sim \mathcal{N}(A\boldsymbol{w} + \hat{\boldsymbol{\mu}}_0^*, \hat{K}^{**}), \tag{17}$$

where $A = H^* - K^* K_v^{-1} H$, H^* is the $(m \times p)$-dimensional matrix with columns $[h_i(x_1^*), \ldots, h_i(x_m^*)]^T$, $i = 1, \ldots, p$, given by the basis function evaluated at the test covariates and $\hat{\boldsymbol{\mu}}_0^* = K^* K_v^{-1}\mathbf{y}$ can be interpreted as the posterior mean of $\mathbf{f}^*$, as it can be derived from equation (5) by setting $\boldsymbol{\mu} = 0$ and $\boldsymbol{\mu}^* = 0$, having assumed $f(x) \sim GP(0, k(x, x'))$.

Then, using (15) and (17) we obtain

$$\begin{aligned}&p(\mathbf{g}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v) \\ &\qquad \propto \int_{\mathcal{W}} p(\mathbf{g}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v, \boldsymbol{w}) p(\boldsymbol{w}|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v) d\boldsymbol{w} \\ &\qquad \propto \int_{\mathcal{W}} p(\mathbf{g}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v, \boldsymbol{w}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_v, \boldsymbol{w}) \exp(\boldsymbol{\ell}^T\boldsymbol{w}) d\boldsymbol{w} \\ &\qquad = \int_{\mathcal{W}} e^{-\frac{1}{2}(\mathbf{g}^* - A\boldsymbol{w} - \hat{\boldsymbol{\mu}}_0^*)^T (\hat{K}^{**})^{-1} (\mathbf{g}^* - A\boldsymbol{w} - \hat{\boldsymbol{\mu}}_0^*) - \frac{1}{2}(\mathbf{y} - H\boldsymbol{w})^T K_v^{-1} (\mathbf{y} - H\boldsymbol{w}) + \boldsymbol{\ell}^T\boldsymbol{w}} d\boldsymbol{w} \\ &\qquad \propto e^{-\frac{1}{2}(\mathbf{g}^* - \hat{\boldsymbol{\mu}}_0^*)^T (\hat{K}^{**})^{-1} (\mathbf{g}^* - \hat{\boldsymbol{\mu}}_0^*) + \frac{1}{2}\overline{\boldsymbol{w}}^T \Sigma^{-1} \overline{\boldsymbol{w}}},\end{aligned}$$

where

$$\Sigma^{-1} = A^T (\hat{K}^{**})^{-1} A + H^T K_v^{-1} H$$

and

$$\overline{\boldsymbol{w}} = \Sigma \left[A^T (\hat{K}^{**})^{-1} (\mathbf{g}^* - \hat{\boldsymbol{\mu}}_0^*) + H^T K_v^{-1}\mathbf{y} + \boldsymbol{\ell} \right].$$

It follows that

$$p(\mathbf{g}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v) \propto \; e^{-\frac{1}{2}(\mathbf{g}^*-\hat{\boldsymbol{\mu}}_g^*)^T(\hat{K}_g^{**})^{-1}(\mathbf{g}^*-\hat{\boldsymbol{\mu}}_g^*)}$$

with

$$(\hat{K}_g^{**})^{-1} = (\hat{K}^{**})^{-1} - (\hat{K}^{**})^{-1}A\Sigma A^T(\hat{K}^{**})^{-1},$$

$$\begin{aligned}\hat{\boldsymbol{\mu}}_g^* =& \hat{K}_g^{**}(\hat{K}^{**})^{-1}\left[(1 - A\Sigma A^T(\hat{K}^{**})^{-1})\hat{\boldsymbol{\mu}}_0^* + A\Sigma\boldsymbol{\ell} + A\Sigma H^T K_v^{-1}\mathbf{y}\right] \\ =& \hat{\boldsymbol{\mu}}_0^* + \hat{K}_g^{**}(\hat{K}^{**})^{-1}A\Sigma\left(\boldsymbol{\ell} + H^T K_v^{-1}\mathbf{y}\right).\end{aligned}$$

By the matrix inversion Lemma (also known as Woodbury formula) one obtains

$$\begin{aligned}\hat{K}_g =& \left[(\hat{K}^{**})^{-1} - (\hat{K}^{**})^{-1}A\Sigma A^T(\hat{K}^{**})^{-1}\right]^{-1} = \hat{K}^{**} + A(H^T K_v^{-1}H)^{-1}A^T = \\ =& K^{**} - K^* K_v^{-1}K^{*T} + (H^* - K^* K_v^{-1}H)(H^T K_v^{-1}H)^{-1}(H^* - K^* K_v^{-1}H)^T.\end{aligned} \tag{18}$$

and

$$\begin{aligned}\Sigma =& \left[H^T K_v^{-1}H + A^T(\hat{K}^{**})^{-1}A\right]^{-1} \\ =& (H^T K_v^{-1}H)^{-1} - (H^T K_v^{-1}H)^{-1}A^T(\hat{K}_g^{**})^{-1}A(H^T K_v^{-1}H)^{-1}.\end{aligned}$$

Then,

$$\begin{aligned}\hat{K}_g^{**}(\hat{K}^{**})^{-1}A\Sigma =& \hat{K}_g^{**}(\hat{K}^{**})^{-1}\left(1 - A(H^T K_v^{-1}H)^{-1}A^T(\hat{K}_g^{**})^{-1}\right)A(H^T K_v^{-1}H)^{-1} \\ =& A(H^T K_v^{-1}H)^{-1},\end{aligned}$$

where we have used the fact that $A(H^T K_v^{-1}H)^{-1}A^T = \hat{K}_g^{**} - \hat{K}^{**}$, and so

$$\hat{\boldsymbol{\mu}}_g^* = \hat{\boldsymbol{\mu}}_0^* + A(H^T K_v^{-1}H)^{-1}\left(\boldsymbol{\ell} + H^T K_v^{-1}\mathbf{y}\right). \tag{19}$$

Proposition 2 then follows from noticing that $\hat{\boldsymbol{\mu}}_g^*$ and $(\hat{K}_g^{**})^{-1}$ correspond to the mean vector and covariance matrix of the GP $g(x)|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v \sim GP(\hat{\mu}_g(x),$ $\hat{k}_g(x, x'))$, with $\hat{\mu}_g(x)$ and $\hat{k}_g(x, x')$ given by (16), evaluated at the test covariates $\mathbf{x}^*$. The conditions $n \geq p$ and $\text{rank}(H) = p$ assure that $H^T K_v^{-1}H$ is invertible. □

When the vectors $h_i(\mathbf{x})$ of some basis functions at the observations covariates are linearly dependent, e.g., $h_i(\mathbf{x}) = \alpha_j h_j(\mathbf{x})$, the condition $\text{rank}(H) = p$

is not verified. This can happen, for instance, if the observations are equally spaced with step s and both h_i and h_j are periodic with period equal to s. Clearly, the value of coefficients w_i, w_j cannot be learned from such data (only the value of $w_i + w_j$ can) and thus the use of an improper prior leads to an improper posterior for w_i and w_j. However, this condition can be easily met by properly choosing the set of basis functions, based on the distribution of the samples covariates.

In Bayesian inference, it is interesting to study convergence properties as the number of the observations increases. As the effect of the prior model on posterior inferences depends on the shape of the base kernel $k_{\boldsymbol{\theta}}(x, x')$ and the distribution of the observations, it is difficult to obtain general convergence results. We therefore study convergence of the model only in the simple case where $k_{\boldsymbol{\theta}}(x, x') = 0$, $\forall x, x'$, that is, when the model in (10) reduces to a linear regression in the space of the basis function $\mathbf{h}(x)$.

2.1.2. Linear regression in the space of the basis functions $\mathbf{h}(x)$

As $k_{\boldsymbol{\theta}}(x, x') = 0$, $\forall x, x'$, the problem in (10) reduces to making inferences about the function $g(x) = \mathbf{h}(x)\boldsymbol{w}$ from n observations affected by white Gaussian noise with known variance σ_v^2.

In this case, $K_v = \sigma_v^2 \mathbf{I}$ and $\mathbf{k}(x) = O_n$ (where O_n a n-dimensional vector of zeros). Using the prior $p(\boldsymbol{w}) \propto \exp(\ell^T \boldsymbol{w})$, we obtain a posterior GP with mean function

$$\hat{\mu}_g(x) = \mathbf{h}(x)(H^T H)^{-1}(H^T \mathbf{y} + \sigma_v^2 \boldsymbol{\ell}), \tag{20}$$

which, for $\boldsymbol{\ell} = 0$ is equivalent to the estimate obtained by least square regression.

Proposition 3. *Let the least square estimate* $\mathbf{h}(x)(H^T H)^{-1} H^T \mathbf{y}$ *converge to the function* $g_{LS}(x)$ *for* $n \to \infty$. *Then, given the model* $g(x) = \mathbf{h}(x)\boldsymbol{w}$ *and the prior* $p(\boldsymbol{w}) \propto \exp(\ell^T \boldsymbol{w})$, *posterior inferences about* $E[g(x)]$ *converge to* $g_{LS}(x)$.

Proof. Let Σ_H be a $(p \times p)$-dimensional matrix with elements ij given by $\overline{h}_{ij} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} h_i(x_k) h_j(x_k)$.Then, the second term of

$$\hat{\mu}_g(x) = \mathbf{h}(x)(H^T H)^{-1} H^T \mathbf{y} + \sigma_v^2 \mathbf{h}(x)(H^T H)^{-1} \boldsymbol{\ell}$$

converges to $\frac{\sigma_v^2}{n} \Sigma_H^{-1} \boldsymbol{\ell} \overset{n\to\infty}{\to} 0$, whereas the first term converges to $g_{LS}(x)$. □

3. An imprecise GP model

The standard objective Bayesian approach using, in case of lack of prior information, *non-informative* priors verifying some desirable invariance properties, does not justify the choice of the improper uniform prior among all possible translation invariant conjugate priors that can be defined. Moreover, a single prior, even when *non-informative*, cannot describe a situation of prior ignorance, as a priori it assumes precise values for all inferences thus modeling a precise prior belief. Improper priors are often considered to be *non-informative* as they lead to posterior inferences that, apparently, are not influenced by the choice of prior parameters. However, we have seen that the uniform prior, traditionally considered *non-informative*, corresponds an improper priors of the form $p(\boldsymbol{w}) \propto \exp(\boldsymbol{\ell}^T \boldsymbol{w})$ with parameter $\boldsymbol{\ell} = O_p$; different choice of the parameter would lead to different posterior inferences. For these reasons, the improper uniform prior cannot be considered to actually model prior ignorance.

Instead, a set of priors providing vacuous prior inferences about some function of interest, better reflects a situation of prior near-ignorance. To build such model, let us consider the set of priors proposed in [7],

$$\mathcal{M} = \{p(\boldsymbol{w}) \propto \exp(\boldsymbol{\ell}^T \boldsymbol{w}),\ \boldsymbol{\ell} \in \mathbb{L}\},$$

where $\mathbb{L}$ is a bounded closed convex subset of $\mathbb{R}^p$ strictly including the origin.

In [7] it has been shown that $\mathcal{M}$ verifies

- translational invariance, as $\underline{E}[\gamma(\boldsymbol{w} - a)] = \underline{E}[\gamma(\boldsymbol{w})]$ holds for any $\boldsymbol{\ell}$ and any bounded function $\gamma(\boldsymbol{w})$ (the same holds for the upper expectation);
- prior ignorance about the expectation of $\boldsymbol{w}$, i.e.,

  $$\inf_{\ell_i \in \mathbb{L}} \underline{E}[w_i] = -\infty, \qquad \sup_{\ell_i \in \mathbb{L}} \overline{E}[w_i] = +\infty,$$

  holds for all $i = 1, \ldots, p$ and for any $\mathbb{L}$. Then, it also holds that

  $$\begin{cases} \underline{E}[w_i] = -\infty, & \forall \ell_i < 0 \\ \overline{E}[w_i] = +\infty, & \forall \ell_i > 0 \end{cases} \tag{21}$$

Definition 3. *Given a covariance kernel $k_{\boldsymbol{\theta}}(x,x')$, a set of basis functions $\mathbf{h}(x)$, and a bounded closed convex subset $\mathbb{L}$ of $\mathbb{R}^p$ strictly including the origin, we define an Imprecise Gaussian Process (IGP) as the set of GPs*

$$\begin{aligned}\mathcal{G} = \{&g(x) = f(x) + \mathbf{h}(x)\boldsymbol{w} : \\ &f(x) \sim GP(0, k_{\boldsymbol{\theta}}(x,x')), p(\boldsymbol{w}) \propto \exp(\boldsymbol{\ell}^T\boldsymbol{w}), \boldsymbol{\ell} \in \mathbb{L}\}.\end{aligned} \tag{22}$$

The set in (22) can be interpreted as the set of GPs

$$\mathcal{G} = \left\{GP(\mathbf{h}(x)\mathbf{b}, k_{\boldsymbol{\theta}}(x,x') + \mathbf{h}(x)B\mathbf{h}(x')^T) : B^{-1}\mathbf{b} = \boldsymbol{\ell} \in \mathbb{L}, B^{-1} \to O_{pp}\right\}.$$

A priori we can state the following result.

Proposition 4. *The IGP model verifies*

- *translation invariance with respect to $g(x)$, i.e., for any bounded real valued function γ*

$$E[\gamma(g(x) - a)] = E[\gamma(g(x))] \ \forall \ a$$

- *prior ignorance about the expectation of $g(x_{p^*+1})$ for any covariate x_{p^*+1} conditional on the value of $E[g(x)] = \mu_g(x)$ at any $p^* < p$ distinct covariates $x_1, \dots, x_{p^*}$, provided that $\mathbf{h}(x_{p^*+1})$ and the set of vectors $\mathbf{h}(x_i)$, $i = 1, \dots, p^*$ are linearly independent, i.e.,*

$$\begin{aligned}&\inf_{\boldsymbol{\ell}\in\mathbb{L}} E[g(x_{p^*+1})|\mu_g(x_1), \dots, \mu_g(x_{p^*})] = -\infty \\ &\sup_{\boldsymbol{\ell}\in\mathbb{L}} E[g(x_{p^*+1})|\mu_g(x_1), \dots, \mu_g(x_{p^*})] = +\infty.\end{aligned} \tag{23}$$

Proof. Concerning translation invariance we have that $E[\gamma(g(x) - a)] = E[\gamma(f(x) + \mathbf{h}(x)\boldsymbol{w} - a)] = E[\gamma(f(x) + \mathbf{h}(x)(\boldsymbol{w} - \mathbf{a}^*))]$, with $\mathbf{a}^* = \left[\frac{a}{h_1(x)}, \dots, \frac{a}{h_p(x)}\right]^T$. Then, from translation invariance of $p(\boldsymbol{w}) \propto \exp(\boldsymbol{\ell}^T\boldsymbol{w})$ if follows that for all priors in the set it holds

$$\underline{E}[\gamma(g(x) - a)] = \underline{E}[\gamma'(\boldsymbol{w} - \mathbf{a}^*)] = \underline{E}[\gamma'(\boldsymbol{w})] = \underline{E}[\gamma(g(x))],$$

where $\gamma'(\boldsymbol{w}) = \gamma(f(x) + \mathbf{h}(x)\boldsymbol{w})$. Then, translation invariance holds also for the prior that attains the inf of $\underline{E}[\gamma]$. The same can be shown for the sup, thus proving translation invariance.

Concerning prior ignorance, let $\mathbf{h}^f = [h_1, \ldots, h_{p^*}]$ and $\mathbf{h}^l = [h_{p^*+1}, \ldots, h_p]$ be the sets of, respectively, the first p^* and the last $p - p^*$ basis functions, and $\boldsymbol{w}_f = [w_1, \ldots, w_{p^*}]^T$ and $\boldsymbol{w}_l = [w_{p^*+1}, \ldots, w_p]^T$ the corresponding coefficient vectors. Moreover, let H_f and H_l be the $(p^* \times p^*)$-dimensional and $(p^* \times (p - p^*))$-dimensional matrices with rows equal to, respectively, $\mathbf{h}^f(x_i)$, and $\mathbf{h}^l(x_i)$, $i = 1, \ldots, p^*$. Knowing the value of $\boldsymbol{\mu}_g = [\mu_g(x_1), \ldots, \mu_g(x_{p^*})]^T$, where $\mu_g(x_i) = \mathbf{h}(x_i)\boldsymbol{w}$, we can write

$$\boldsymbol{\mu}_g = H_f \boldsymbol{w}_f + H_l \boldsymbol{w}_l \implies \boldsymbol{w}_f = H_f^{-1}(\boldsymbol{\mu}_g - H_l \boldsymbol{w}_l).$$

Then,

$$\begin{aligned} E[g(x_{p^*+1})] =& \mathbf{h}^f(x_{p^*+1})\boldsymbol{w}_f + \mathbf{h}^l(x_{p^*+1})\boldsymbol{w}_l \\ =& \mathbf{h}^f(x_{p^*+1})H_f^{-1}(\boldsymbol{\mu}_g - H_l\boldsymbol{w}_l) + \mathbf{h}^l(x_{p^*+1})\boldsymbol{w}_l \\ =& \mathbf{h}^f(x_{p^*+1})H_f^{-1}\boldsymbol{\mu}_g - (\mathbf{h}^f(x_{p^*+1})H_f^{-1}H_l - \mathbf{h}^l(x_{p^*+1}))\boldsymbol{w}_l \end{aligned} \tag{24}$$

where $\mathbf{h}^f(x_{p^*+1})H_f^{-1}H_l - \mathbf{h}^l(x_{p^*+1}) \neq 0$ as $\mathbf{h}^l(x_{p^*+1})$ and the rows of H_l are linearly independent. It follows from (21) that

$$\begin{aligned} &\inf_{\boldsymbol{\ell} \in \mathbb{L}} \left(\mathbf{h}^f(x_{p^*+1})H_f^{-1}\boldsymbol{\mu}_g - (\mathbf{h}^f(x_{p^*+1})H_f^{-1}H_l - \mathbf{h}^l(x_{p^*+1}))\boldsymbol{w}_l\right) = -\infty, \\ &\sup_{\boldsymbol{\ell} \in \mathbb{L}} \left(\mathbf{h}^f(x_{p^*+1})H_f^{-1}\boldsymbol{\mu}_g - (\mathbf{h}^f(x_{p^*+1})H_f^{-1}H_l - \mathbf{h}^l(x_{p^*+1}))\boldsymbol{w}_l\right) = +\infty, \end{aligned}$$

where the inf is found by choosing for all $i = p^* + 1, \ldots, p$

$$\begin{cases} \ell_i < 0 & \text{if } (\mathbf{h}^f(x^*)H_f^{-1}h_i(\mathbf{x}) - h_i(x^*)) < 0, \\ \ell_i > 0 & \text{if } (\mathbf{h}^f(x^*)H_f^{-1}h_i(\mathbf{x}) - h_i(x^*)) > 0, \end{cases}$$

and vice versa for the sup. □

Proposition 4 states that an IGP with p basis functions models a sort of "uncorrelated" prior ignorance about the expectation of the vector $\mathbf{g}$ of the values of $g(x)$ evaluated at any p distinct covariate values, in the sense that prior inferences about $E[g(x_i)]$ remain vacuous independently of the value of $E[g(x_j)]$ at the other $p - 1$ covariates $x_j \neq x_i$. On the other side, the expectation of $g(x_{p+1})$ at a new distinct covariate x_{p+1} is fully determined by the knowledge of the expectation of $g(x)$ at p different covariate values (provided that the vectors $\mathbf{h}(x_i)$, $i = 1, \ldots, p$ are linearly independent) and is, indeed, equal to $E[g(x_{p+1})|\boldsymbol{\mu}_g] = \mathbf{h}(x_{p+1})H^{-1}\boldsymbol{\mu}_g$, with $\boldsymbol{\mu}_g = [\mu_g(x_1), \ldots, \mu_g(x_p)]^T$.

As discussed in Section 2, MAP estimates of the hyperparameters $\boldsymbol{\theta}_v$ are used in the model. However the different priors in the IGP set produce different estimates, whereas the IGP model here proposed requires the same set of hyperparameters for all priors. The issue is, then, which of the IGP priors should be used to estimate $\boldsymbol{\theta}_v$. As all IGP models include, by definition, the limit exponential prior with $\boldsymbol{\ell} = O_p$, we suggest to use as set of hyperparameters $\boldsymbol{\theta}_v$ the MAP estimates derived in this case, i.e., those obtained by maximizing $L(\boldsymbol{\theta}_v) + p(\boldsymbol{\theta}_v)$, where $L(\boldsymbol{\theta}_v)$ is given by (14) when $\boldsymbol{\ell} = O_p$.

Concerning posterior inferences we can state the following result.

Proposition 5. *Under the IGP model, the bounds of the posterior expectation of $g(x)$ are*

$$\begin{aligned} \underline{\mu}_g(x) &= \inf_{\boldsymbol{\ell}\in\mathbb{L}} E[g(x)|\mathbf{x},\mathbf{y},\boldsymbol{\theta}_v] = \hat{\mu}_0(x) + \underline{\mu}_1(x) \\ \overline{\mu}_g(x) &= \sup_{\boldsymbol{\ell}\in\mathbb{L}} E[g(x)|\mathbf{x},\mathbf{y},\boldsymbol{\theta}_v] = \hat{\mu}_0(x) + \overline{\mu}_1(x), \end{aligned} \tag{25}$$

where

$$\hat{\mu}_0(x) = \mathbf{k}(x)^T K_v^{-1}\mathbf{y} + (\mathbf{h}(x)^T - H^T K_v^{-1}\mathbf{k}(x))^T (H^T K_v^{-1} H)^{-1} H^T K_v^{-1}\mathbf{y}$$

and $\underline{\mu}_\ell(x)$ and $\overline{\mu}_\ell(x)$ are found by solving two convex optimization problems

$$\underline{\mu}_1(x) = \min_{\boldsymbol{\ell}\in\mathbb{L}} \left[\hat{\boldsymbol{\mu}}_1(x)^T\boldsymbol{\ell}\right], \qquad \overline{\mu}_1(x) = \max_{\boldsymbol{\ell}\in\mathbb{L}} \left[\hat{\boldsymbol{\mu}}_1(x)^T\boldsymbol{\ell}\right], \tag{26}$$

where $\hat{\boldsymbol{\mu}}_1(x)^T = (\mathbf{h}(x)^T - H^T K_v^{-1}\mathbf{k}(x))^T (H^T K_v^{-1} H)^{-1}$.

Proof. As the lower and upper bounds or $E[g(x)|\mathbf{x},\mathbf{y}]$ are found by minimizing and, respectively, maximizing $\hat{\mu}_g(x)$, this result follows directly from (16) by rewriting it as $\hat{\mu}_g(x) = \hat{\mu}_0(x) + \hat{\boldsymbol{\mu}}_1(x)\boldsymbol{\ell}$. □

Notice that if $\mathbb{L}$ is a polytope, the two optimizations in proposition 5 reduce to a linear programming problem. Moreover, by assuming a regular shape for $\mathbb{L}$ we can obtain posterior inferences in a closed form.

Proposition 6. *Given an IGP model with $\mathbb{L}$ defined as*

1. *the hyperrectangle $\mathbb{L} = [\underline{\ell}_1, \overline{\ell}_1] \times [\underline{\ell}_2, \overline{\ell}_2] \times \cdots \times [\underline{\ell}_m, \overline{\ell}_m]$, with $\underline{\ell}_i < 0$ and $\overline{\ell}_i > 0$ for all $i = 1, \ldots, p$.*
2. *the hypercube $\mathbb{L} = \{\boldsymbol{\ell} : ||\boldsymbol{\ell}||_\infty \leq c\}$*

3. *the cross-polytope* $\mathbb{L} = \{\boldsymbol{\ell} : ||\boldsymbol{\ell}||_1 \leq c\}$,

where $||\mathbf{v}||_q$ *is the q-norm of a vector* $\mathbf{v}$, *the value of* $\underline{\mu}_1(x)$ *and* $\overline{\mu}_1(x)$ *in (26) are, respectively,*

$$\begin{array}{lll} \textit{1.} & \underline{\mu}_1(x) = \hat{\boldsymbol{\mu}}_1^T(x)\underline{\boldsymbol{\ell}}^* & \overline{\mu}_1(x) = \hat{\boldsymbol{\mu}}_1^T(x)\overline{\boldsymbol{\ell}}^*; \\ \textit{2.} & \underline{\mu}_1(x) = -c||\hat{\boldsymbol{\mu}}_1^T(x)||_1 & \overline{\mu}_1(x) = c||\hat{\boldsymbol{\mu}}_1^T(x)||_1; \\ \textit{3.} & \underline{\mu}_1(x) = -c||\hat{\boldsymbol{\mu}}_1(x)||_\infty & \overline{\mu}_1(x) = c||\hat{\boldsymbol{\mu}}_1(x)||_\infty. \end{array}$$

where $\underline{\boldsymbol{\ell}}^* = [\underline{\ell}_1^*, \ldots, \underline{\ell}_p^*]^T$ *is obtained as*

$$\underline{\ell}_i^* = \begin{cases} \underline{\ell}_i & \textit{if } [\hat{\boldsymbol{\mu}}_1(x)]_i \geq 0, \\ \overline{\ell}_i & \textit{if } [\hat{\boldsymbol{\mu}}_1(x)]_i < 0, \end{cases}$$

and vice versa for $\overline{\boldsymbol{\ell}}^* = [\overline{\ell}_1^*, \ldots, \overline{\ell}_p^*]^T$.

Proof. From (26), $\underline{\mu}_1(x)$ and $\overline{\mu}_1(x)$ are found by minimizing the summation

$$\hat{\boldsymbol{\mu}}_1^T(x)\boldsymbol{\ell} = \sum_{i=1} \hat{\mu}_i^1(x)\ell_i, \tag{27}$$

where $\hat{\mu}_i^1(x)$ is the i-th element of the vector $\hat{\boldsymbol{\mu}}_1(x)$.
The results at points 1. and 2. follow from the fact that each term of (27) can be optimized independently from the others and it monotonically increases with ℓ_i if $\hat{\mu}_i^1(x) > 0$ and decreases if $\hat{\mu}_i^1(x) < 0$.
The result at point 3. is found by considering that the vertexes of a cross-polytope are represented by vectors $[\ell_1, \ldots, \ell_p]$ with a single element $\ell_i = \pm c$ and all other elements ℓ_j, $j \neq i$, equal to 0. Then, the bounds of $\hat{\boldsymbol{\mu}}_1^T(x)\boldsymbol{\ell}$ are found at the vertexes that assign values c and $-c$ to the coefficient ℓ_i of the element $\hat{\mu}_i^1(x)$ with the maximum absolute value. □

From propositions 5 and 6 it follows that the imprecision in the three cases considered is equal to, respectively,

$$\begin{aligned} &1.\ \overline{\mu}_g(x) - \underline{\mu}_g(x) = ||\hat{\boldsymbol{\mu}}_1^T(x)(\overline{\boldsymbol{\ell}} - \underline{\boldsymbol{\ell}})||_1; \\ &2.\ \overline{\mu}_g(x) - \underline{\mu}_g(x) = 2c||\hat{\boldsymbol{\mu}}_1||_1; \\ &3.\ \overline{\mu}_g(x) - \underline{\mu}_g(x) = 2c||\hat{\boldsymbol{\mu}}_1||_\infty, \end{aligned} \tag{28}$$

where $\underline{\boldsymbol{\ell}} = [\underline{\ell}_1, \ldots, \underline{\ell}_p]^T$ and $\overline{\boldsymbol{\ell}} = [\overline{\ell}_1, \ldots, \overline{\ell}_p]^T$. One can see that, in all three cases, the set $\mathbb{L}$ (and therefore parameter c for the hypercube and the cross-polytope) controls the degree of imprecision of posterior inferences. Moreover, the posterior imprecision does not depend on $\mathbf{y}$ and thus it is invariant with respect to re-parametrizations of the function space.

The hypercube and the hyperrectagle choices for the shape of $\mathbb{L}$ model a sort of independence between the parameters in $\boldsymbol{\ell}$, whereas the cross-polytope introduces dependences among parameter (e.g., if $\ell_1 = c$, then $\ell_j = 0 \ \forall j = 2, \ldots, p$). Moreover, the hypercube, compared to hyperrectangle, requires the elicitation of a single parameter c. Therefore, as in the lack of prior knowlegde we aim to reduce the amount of information injected into the model by prior assumptions, we will hereafter focus on the hypercube shaped set $\mathbb{L} = \{\boldsymbol{\ell} : ||\boldsymbol{\ell}||_\infty < c\}$ and call the corresponding model c-IGP.

As already discussed in section 2.1.1, posterior inferences depend on the base kernel and the distribution of the observation covariates, and thus it is difficult to obtain general convergence results. However, convergence can be proven under some assumption.

Proposition 7. *Given a set of training data with covariates x_i, $i = 1, \ldots, n$ and a test covariate x^* such that*

- $k_{\boldsymbol{\theta}}(x_i, x_j) = 0$ *for all* $i, j = 1, \ldots, n$,
- $k_{\boldsymbol{\theta}}(x^*, x_i) \neq 0$ *for at most* n^* *training covariates* x_i, *with* $\lim\limits_{n\to\infty} \frac{n^*}{n} = 0$

it holds

$$\lim_{n\to\infty} \overline{\mu}_g(x^*) - \underline{\mu}_g(x^*) = 0.$$

Proof. From the first assumption we can write $K_v = \sigma_v^2 \mathbf{I}$. Then from (28)

$$\overline{\mu}_g(x^*) - \underline{\mu}_g(x^*) = 2c \left| \left(\mathbf{h}(x)^T - \frac{H^T \mathbf{k}(x)}{\sigma_v^2} \right)^T \sigma_v^2 (H^T H)^{-1} \right| \mathbb{1}_p. \tag{29}$$

where $\mathbb{1}_p$ is a p-dimensional vector of ones.From the following inequality,

$$|h_i(\mathbf{x})\mathbf{k}(x^*)| = \left| \sum_{l : k_{\boldsymbol{\theta}}(x^*, x_l) \neq 0} k_{\boldsymbol{\theta}}(x^*, x_l) h_i(x_l) \right| \leq n^* k^{max}(x^*) h_i^{max}$$

where $k^{max}(x^*) = \max\limits_{l:k_{\boldsymbol{\theta}}(x^*,x_l)\neq 0} |k_{\boldsymbol{\theta}}(x^*,x_l)|$ and $h_i^{max} = \max\limits_{l:k_{\boldsymbol{\theta}}(x^*,x_l)\neq 0} |h_i(x_l)|$ we can write

$$\overline{\mu}_g(x^*) - \underline{\mu}_g(x^*) \leq \frac{2c}{\sigma_v^2}\left(|\mathbf{h}(x)^T| + \frac{n^* k^{max}(x^*)\mathbf{h}^{max}}{\sigma_v^2}\right)^T |(H^T H)^{-1}|\mathbb{1}_p,$$

where $\mathbf{h}^{max} = [h_1^{max}, \ldots, h_p^{max}]^T$. Having defined Σ_H^{-1} as in the proof of Proposition 7 we can write

$$\begin{aligned} &\lim_{n\to\infty} \frac{2c}{\sigma_v^2}\left(|\mathbf{h}(x)| + \frac{n^* k^{max}(x^*)\mathbf{h}^{max}}{\sigma_v^2}\right)^T |(H^T H)^{-1}|\mathbb{1}_p \\ &= \lim_{n\to\infty} \frac{2c}{\sigma_v^2 n}\left(|\mathbf{h}(x)| + \frac{n^* k^{max}(x^*)\mathbf{h}^{max}}{\sigma_v^2}\right)^T |\Sigma_H^{-1}|\mathbb{1}_p = 0. \end{aligned}$$

Therefore, also $\overline{\mu}_g(x^*) - \underline{\mu}_g(x^*)$ converges to 0. □

Proposition 7 implies that in case of linear regression in the space of the basis functions, the imprecise model converges to the precise model with improper uniform prior, thus verifying the convergence property proposed in [7] as one of the main requirements for a model of prior near-ignorance. Notice that this result is a natural consequence of Proposition 3, as for all priors in the set $\mathcal{M}$, the posterior expectation converges to the least square estimate.

Although the conditions in proposition 7 are very restrictive, when using a kernel that goes quickly to zero after a certain distance, like, for instance, the exponential kernel, and the training covariates span a range much larger than the distance of influence of the kernel (lengthscale), the number of (almost) uncorrelated pairs of observations will increase with n, whereas only a bounded number of them will be (significantly) correlated with the test covariate x^*. Although different, this situation will be similar to the one described in Proposition 7. Therefore, we conjecture that when using a covariance function whose influence cancels out beyond a bounded distance, like with the exponential kernel, convergence can be satisfied provided that the observations span an interval much larger than the kernel lengthscale.

It is often useful to compute pointwise credible intervals $CI_\alpha(x) = [\underline{g}_\alpha(x), \overline{g}_\alpha(x)]$ for the value of $g(x)$. Using a GP prior $g(x) \sim GP(\mu(x),\, k_{\boldsymbol{\theta}}(x,x'))$, a posterior $(1-\alpha)\%$ credible interval for the value of $g(x)$ is $CI_\alpha(x) = [\hat{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x,x)}, \hat{\mu}(x) + z_{\alpha/2}\sqrt{\hat{k}(x,x)}]$ where $\hat{\mu}(x)$ and $\hat{k}(x,x')$ are given in (7) and $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. Hence we have that the posterior probability $P\big(g(x) \in CI_\alpha(x)\big)$ is $1-\alpha$. In the imprecise case, we define the credible interval by imposing that the upper posterior probabilities $\overline{P}\big(g(x) < \underline{g}_\alpha(x)\big)$ and $\overline{P}\big(g(x) > \overline{g}_\alpha(x)\big)$ are equal to $\alpha/2$. This implies that $P\big(g(x) \in CI_\alpha(x)\big) \geq 1-\alpha$ for all GPs in $\mathcal{G}$.

Proposition 8. *Under the c-IGP model, the interval*

$$CI_\alpha(x) = \left[\underline{g}_\alpha(x) = \underline{\mu}_g(x) - z_{\alpha/2}\sqrt{\hat{k}_g(x,x)}, \overline{g}_\alpha(x) = \overline{\mu}_g(x) + z_{\alpha/2}\sqrt{\hat{k}_g(x,x)}\right],$$

where $\underline{\mu}_g(x)$, $\overline{\mu}_g(x)$ and $\hat{k}_g(x,x')$ are given in (25) and (16), verifies

$$\overline{P}\big(g(x) < \underline{g}_\alpha(x)\big) = \alpha/2, \quad \overline{P}\big(g(x) > \overline{g}_\alpha(x)\big) = \alpha/2$$

Proof. For each GP in $\mathcal{G}$ it holds that the lower bound of a $(1-\alpha)\%$ credible interval is $\hat{\mu}_g(x) - z_{\alpha/2}\sqrt{\hat{k}_g(x,x)}$, where $\hat{\mu}_g(x)$ and $\hat{k}_g(x,x')$ are given (16). Then

$$P(g(x) < \hat{\mu}_g(x) - z_{\alpha/2}\sqrt{\hat{k}_g(x,x)}) = \alpha/2,$$

and thus

$$\begin{aligned} P&\left(g(x) < \min_{\boldsymbol{\ell}\in\mathbb{L}} \hat{\mu}_g(x) - z_{\alpha/2}\sqrt{\hat{k}_g(x,x)}\right) \\ &= P\left(g(x) < \underline{\mu}_g(x) - z_{\alpha/2}\sqrt{\hat{k}_g(x,x)}\right) \leq \alpha/2. \end{aligned}$$

Similarly we can prove that

$$P\left(g(x) > \overline{\mu}(x) + z_{\alpha/2}\sqrt{\hat{k}_g(x,x)}\right) \leq \alpha/2.$$

□

As $\hat{k}_g(x,x)$ is the same in the precise and imprecise models, the width of the pointwise CIs increases in the imprecise model only as much as the difference $\overline{\mu}_g(x) - \underline{\mu}_g(x)$ between the upper and lower bounds of $E[g(x)]$.

Data analyst are often interested also in simultaneous credible regions (SCR) for the vector of values of $g(x)$ at multiple covariates values $\mathbf{x}^*$. Given a GP $g(x) \sim GP(\mu(x), k_{\boldsymbol{\theta}}(x, x'))$ and vector of m covariate values $\mathbf{x}^*$, a $(1-\alpha)$-SCR for $\mathbf{g}^*$, hereafter denoted as SCR_α, includes all m-dimensional vectors $\hat{\mathbf{g}}^*$ that verify

$$\chi^2(\hat{\mathbf{g}}^*) = (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}^*)^T (\hat{K}^{**})^{-1} (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}^*) < \chi^{-1}(1-\alpha|m), \tag{30}$$

where $\hat{\boldsymbol{\mu}}^*$ and $\hat{K}^{**}$ are given in, respectively, (5) and (6) and $\chi^{-1}(1-\alpha|m)$ is the $(1-\alpha)$-quantile of a Chi-squared distribution with m degrees of freedoms. This implies that the $(1-\alpha)$-SCR verifies $P(\mathbf{g}^* \in SCR_\alpha) = 1-\alpha$. In the imprecise case, we define the posterior SCR as the subset of $\mathbb{R}^m$ such that $\underline{P}(\mathbf{g}^* \in SCR_\alpha) = 1-\alpha$.

Proposition 9. *Given the c-IGP model, the region SCR_α including all m-dimensional vectors $\hat{\mathbf{g}}^*$ such that*

$$\underline{\chi}^2(\hat{\mathbf{g}}^*) = \inf_{\boldsymbol{\ell} \in \mathbb{L}} \chi^2(\hat{\mathbf{g}}^*) < \chi^{-1}(1-\alpha|m),$$

where

$$\chi^2(\hat{\mathbf{g}}^*) = (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}_g^*)^T (\hat{K}_g^{**})^{-1} (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}_g^*) < \chi^{-1}(1-\alpha|m)$$

and $\hat{\boldsymbol{\mu}}_g^*$ and $\hat{K}_g^{**}$ are given in, respectively, (19) and (18), verifies $\underline{P}(\mathbf{g}^* \in SCR_\alpha) = 1-\alpha$.

Proof. As the lower bound of $\chi^2(\hat{\mathbf{g}}^*)$ verifies $\underline{\chi}^2(\hat{\mathbf{g}}^*) < \chi^{-1}(1-\alpha|m)$, all $\hat{\mathbf{g}}^* \in SCR_\alpha$ verify $\chi^2(\hat{\mathbf{g}}^*) < \chi^{-1}(1-\alpha^*|m)$, with $\alpha^* \leq \alpha$, and thus $P(\mathbf{g}^* \in SCR_\alpha) \geq 1-\alpha$. □

The lower bound of $\chi^2(\mathbf{g}^*)$ is found by solving numerically the optimization problem

$$\min_{\boldsymbol{\ell} \in \mathbb{L}} (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}_0^* - \hat{\boldsymbol{\mu}}_1^* \boldsymbol{\ell})^T (\hat{K}_g^{**})^{-1} (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}_0^* - \hat{\boldsymbol{\mu}}_1^* \boldsymbol{\ell}), \tag{31}$$

where $\hat{\boldsymbol{\mu}}_0^* = [\hat{\mu}_0(x_1^*), \ldots, \hat{\mu}_0(x_m^*)]^T$ and $\hat{\boldsymbol{\mu}}_1^*$ is a $(m \times p)$-dimensional matrix with rows equal to $\hat{\boldsymbol{\mu}}_1(x_j^*)$, $j = 1, \ldots, m$. Notice that this is a quadratic optimization problem. In fact, (31) can be rewritten as

$$(\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}_0^*)^T (\hat{K}_g^{**})^{-1} (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}_0^*) + \min_{\boldsymbol{\ell} \in \mathbb{L}} \left(\boldsymbol{\ell}^T Q \boldsymbol{\ell} + \mathbf{v}^T \boldsymbol{\ell} \right),$$

with $Q = \hat{\boldsymbol{\mu}}_1^{*T} (\hat{K}_g^{**})^{-1} \hat{\boldsymbol{\mu}}_1^*$ and $\mathbf{v} = -2\hat{\boldsymbol{\mu}}_1^{*T} (\hat{K}_g^{**})^{-1} (\hat{\mathbf{g}}^* - \hat{\boldsymbol{\mu}}_0^*)$.

In the remaining of this subsection, we provide two examples of IGP models with some illustrative numerical results.

3.1. Imprecise GP with constant mean function

Consider a IGP model with a single constant basis function $h(x) = 1$. This corresponds to a set of prior GPs with constant mean function $\mathbf{b}$ free to vary from $-\infty$ to $+\infty$.

$$\mathcal{G}_c = \left\{GP(b, k_{\boldsymbol{\theta}}(x,x') + B),\ B^{-1}b = \ell \in [-c,c],\ B^{-1} \to 0\right\}. \tag{32}$$

This model has strong analogies with the one presented in [14] which considers the set of priors

$$\begin{aligned}\mathcal{G}_c =& \left\{GP\left(Mh, k_{\boldsymbol{\theta}}(x,x') + \frac{M+1}{c}\right) :\ h = \pm 1,\ M \geq 0\right\}\\ =& \left\{GP\left(M, k_{\boldsymbol{\theta}}(x,x') + \frac{|M|+1}{c}\right) : M \in \mathbb{R}\right\}\\ =& \left\{GP\left(b, k_{\boldsymbol{\theta}}(x,x') + B\right) : b = M, B^{-1}b = c\frac{M}{1+|M|} \in [-c,c]\right\}.\end{aligned}$$

Thus, (32) differs from the model in [14] because it includes only those priors with $M \to \infty$ and does not assume $b = M$ but $b \in \mathbb{R}$.

As discussed above, the constant parameter c determines the degree of posterior imprecision. However imprecision also depend on the base kernel and the training covariates. From (29) the posterior imprecision at a test covariate x^* is

$$\begin{aligned}\overline{\mu}(x^*) - \underline{\mu}(x^*) =& 2c\left|\left(\mathbf{h}(x^*) - H^T K_v^{-1}\mathbf{k}(x^*)\right)^T (H^T K_v^{-1} H)^{-1}\right|\\ =& 2c\left|\frac{1 - \mathbb{1}_n^T K_v^{-1}\mathbf{k}(x^*)}{\mathbb{1}_n^T K_v^{-1}\mathbb{1}_n}\right|.\end{aligned} \tag{33}$$

It follows that if $k_{\boldsymbol{\theta}}(x,x') = 0$ for all x, x', then $K_v = \sigma_v^2\mathbf{I}$ and thus

$$\overline{\mu}(x^*) - \underline{\mu}(x^*) = 2c\frac{\sigma_v^2}{n}.$$

This result is equivalent to the one shown in Section 4.2 of [7], as this choice of the kernel and set of basis functions correspond to the constant model $g(x) = w$, and thus the GP regression reduces to the problem of estimating the mean of a normally distributed variable with known variance σ_v^2 from n observations. As expected from proposition 7, the imprecision goes to zero for $n \to \infty$, and the lower and upper bounds of $E[g(x)]$ converge to the inferences obtained from a GP model using the (improper) uniform prior ($\ell = 0$) on w, thus verifying the convergence property (A4) in [7].

Notice however that the IGP model does not necessarily converge as $n \to \infty$. Consider the following illustrative example.

Example 1. *Assume that the kernel base function $k_{\boldsymbol{\theta}}(x_i, x_j)$ is constantly equal to σ_b^2 for all $i, j = 1, \ldots, n$, whereas $k_{\boldsymbol{\theta}}(x^*, x_i) = 0$ for all x_i, $i = 1, \ldots, n$. Then, $K_v^{-1} = \sigma_v^2 \mathbf{I} + \sigma_b^2 \mathbb{1}_n \mathbb{1}_n^T$ and thus from (33) we have*

$$\begin{aligned}\overline{\mu}(x^*) - \underline{\mu}(x^*) =& 2c \left| \frac{1}{\mathbb{1}_n^T (\sigma_v^2 \mathbf{I} + \sigma_b^2 \mathbb{1}_n \mathbb{1}_n^T)^{-1} \mathbb{1}_n} \right| \\ =& 2c \frac{\sigma_v^2 + n\sigma_b^2}{n} \overset{n \to \infty}{\to} 2c\sigma_b^2.\end{aligned}$$

where we have used the Sherman-Morrison formula to derive the inverse of K_v.

This example describes, for instance, a situation where all training data are collected at the same covariate $x_0 = x_1, x_2, \ldots, x_n$ and $g(x_0)$ is uncorrelated to $g(x^*)$. As the number of observation increases, one will increase the knowledge about $g(x_0)$. However, even knowing the exact value of $g(x_0)$ does not convey a precise knowledge about the uncorrelated value of $g(x^*)$.

Let us consider now a numerical example where the IGP with constant mean is used to make inferences about an unknown underlying regression function.

Example 2. *A sample of $n = 50$ observations affected by white Gaussian noise with $\sigma_v = 0.1$ is drawn from the function $g(x) = \exp(-x^2)$. The covariates $x_1, \ldots, x_n$ are uniformly distributed in $[-1.5, 1.5]$, i.e., $x \sim U[-1.5, 1.5]$. The function is modeled by the precise GP model in (10) with improper uniform prior and the IGP model with constant mean. We choose the squared-exponential kernel in (8) as base kernel $k_{\boldsymbol{\theta}}$. Figure 1 shows the posterior expectation of the GP and the upper and lower expectations of the IGP for different values of c (left) and compares the credible intervals of the precise model and the IGP with $c = 10$ (right). Notice that in the region where there are observations ($x \in [-1, 1]$) the imprecision remains very small even when c is large, whereas it increases significantly outside this region. The same happens for the width of the CIs.*

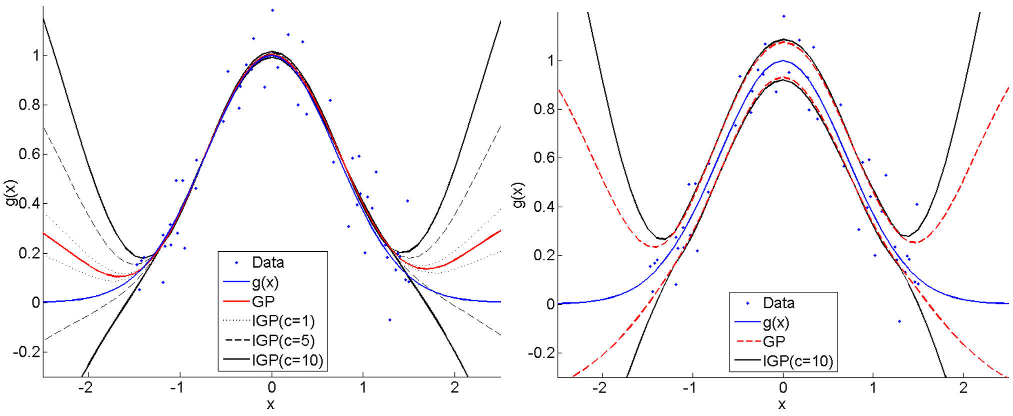

Figure 1: GP and IGP estimates of the function $g(x)$ (left) and pointwise credible intervals (right) given $n = 50$ observations.

3.1.1. IGP polynomial regression

In polynomial regression the set of basis functions $\mathbf{h}(x) = (x^{p-1}, x^{p-2}, \ldots, x, 1)$ is used. Figure 2 shows the posterior expectation for the IGP with $c = 5$ when the mean is modeled by the set of basis functions of the polynomial regression for different choices of p. Again, we can see that imprecision increases very quickly outside the region where data are observed, and in this case, contrarily to the IGP with constant mean, can go to infinity with the distance $|x^* - x_i|$ between the test and training covariates. To show this, let us consider once again the case $k_{\boldsymbol{\theta}}(x, x') = 0$, and assume $p = 2$. Then, from (29) we derive

$$\begin{aligned}
\overline{\mu}(x^*) - \underline{\mu}(x^*) =& \frac{2c\sigma_n^2}{n} \left| \begin{bmatrix} 1 & x^* \end{bmatrix} \begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right| \\
=& \frac{2c\sigma_n^2}{n(\overline{x^2} - \overline{x}^2)} \left| x^*(1 - \overline{x}) + \overline{x^2} - \overline{x} \right| \xrightarrow{x^* \to \infty} \infty,
\end{aligned}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$. This result is reasonable as it reflects the fact that our knowledge about the mean of a polynomial regression function at x^* decreases as the distance of x^* from the training data increases.

From Figure 2 we can also observe that the imprecision increases with the cardinality p of the set of basis functions. However, the IGP polynomial

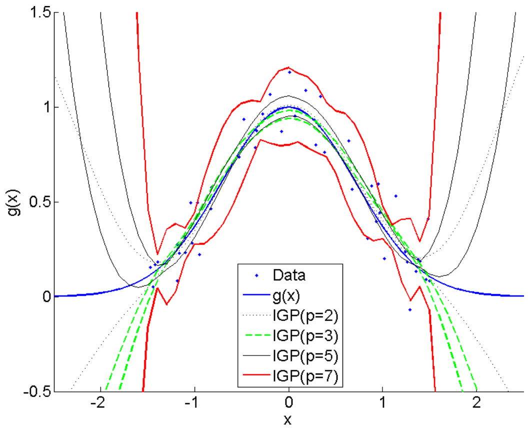

Figure 2: IGP estimates of the function $g(x)$ given $n = 50$ observations using $c = 5$.

regression does not verifies the desirable property that posterior inferences obtained for p always encompass those obtained with $p^* < p$. In fact, it is well known that the choice of the covariance function, and analogously of the set of basis functions, is critical when using a GP, especially when the function at the test covariate x^* has small correlation with the training observations at $\mathbf{x}$, because such observations can only provide little information about the value of the function at x^*, and thus prior information is determinant. To select a model for the correlation, it is often suggested to adopt the choice that maximizes the marginal likelihood, but, as already pointed out at the end of Proposition 1, this is in conflict with the choice of an improper prior, as the marginal likelihood is not bounded in this case.

Due to this limitation of the GPs, in the next section we present some preliminary work to model the lack of prior knowledge about the correlation function by using sets of basis functions.

3.2. An IGP model with imprecise basis functions

As seen in the example of Figure 2, the IGP model cannot always catch the actual lack of prior knowledge, as the precise definition of a set of basis functions, introduces strong prior information. The consequence is that the choice of the set of basis functions can affect significantly posterior inferences. Therefore, such choice should rely on well grounded assumptions about the shape of $g(x)$. However, in the lack of prior knowledge, this is not always possible. In this case, it is more appropriate to account for this state of ignorance, by letting the basis functions $h_i(x)$ free to vary in some set of functions $\mathcal{H}_i$. In this Section, this approach is presented using a set of basis function of cardinality $p = 1$, but we aim to extend it to other settings in future work.

As for the IGP model, we consider a set of p basis functions $\mathbf{h}(x)$, but in this case we combine them based on a vector of fixed parameters $\mathbf{a} = [a_1, \dots, a_{p-1}, 1] = [\mathbf{a}^-, 1]$, thus obtaining the function $h^{IB}(x, \mathbf{a}) = \mathbf{h}(x)\mathbf{a}$. We can then consider the model

$$g(x) = f(x) + wh^{IB}(x, \mathbf{a}) = f(x) + w\mathbf{h}(x)\mathbf{a}$$

and assume for the parameter w the set of priors $\mathcal{M} = \{p(w) \propto \exp(\ell w),\ \ell \in [-c, c]\}$. This corresponds to the IGP model with single basis function $h^{IB}(x, \mathbf{a})$, i.e.,

$$\begin{aligned}\mathcal{G}(\mathbf{a}) = \big\{g(x) = f(x) + wh^{IB}(x, \mathbf{a}) :\ & f(x) \sim GP(0, k_{\boldsymbol{\theta}}(x, x')),\\ & p(w) \propto \exp(\ell w),\ \ell \in [-c, c]\big\}\,.\end{aligned}$$

Definition 4. *Given a set of basis functions $\mathbf{h}(x)$, and a base kernel $k_{\boldsymbol{\theta}}(x, x')$, we define an IGP with imprecise basis function (IGP-IB) the set of IGPs:*

$$\mathcal{G}_{IB} = \big\{\mathcal{G}(\mathbf{a}) : \mathbf{a} = [\mathbf{a}^-, 1], \mathbf{a}^- \in \mathbb{R}^{p-1}\big\}. \tag{34}$$

Proposition 10. *The IGP-IB model verifies prior ignorance as defined in equation (23), for any x_{p^*+1} such that $\mathbf{h}(x_{p^*+1})$ and the set of vectors $\mathbf{h}(x_i)$, $i = 1, \dots, p^*$ are linearly independent.*

Proof. Prior ignorance can be proven in a similar way as for proposition 4, by taking $\boldsymbol{w}_f = w[a_1, \dots, a_{p^*}]^T$ and $\boldsymbol{w}_l = w[a_{p^*}, \dots, a_{p-1}, 1]^T$. Then, from (24) it follows that

$$\inf_{\ell \in [-c,c]} \underline{E}[g(x^*)] = -\infty, \qquad \sup_{\ell \in [-c,c]} \overline{E}[g(x^*)] = +\infty.$$

where the inf is found by choosing for all $i = p^* + 1, \dots, p$

$$\begin{cases} \ell a_i < 0 & \text{if } (\mathbf{h}^f(x^*)H_f^{-1}h_i(\mathbf{x}) - h_i(x^*)) < 0, \\ \ell a_i > 0 & \text{if } (\mathbf{h}^f(x^*)H_f^{-1}h_i(\mathbf{x}) - h_i(x^*)) > 0, \end{cases}$$

with $a_p = 1$, and vice versa for the sup. □

Again, one has to select a set of hyperparameters $\boldsymbol{\theta}_v$. This choice has to be reasonable for all basis functions in the set $\mathcal{G}_{IB}$. If parameters are obtained as MAP estimates from a generic model with $a_i \neq 0$ for all $i = 1, \dots, p-1$, then, the resulting GP $f(x)$ will model the residuals of a linear regression in the transformed covariate $h^{IB}(x, \mathbf{a})$ and thus we can expect it to be a poor model for different values of $\mathbf{a}^-$ (consider for instance the extreme case where all coefficients $a_i = 0$, $i = 1, \dots, p-1$, i.e., $h^{IB}(x, \mathbf{a}) = 1$). Therefore, we suggest to adopt the parameters $\boldsymbol{\theta}_v$ that optimize the model in the case $h^{IB}(x, \mathbf{a}) = 1$.

Proposition 11. *Given the IGP-IB model, the upper and lower posterior bounds of $E[g(x)]$ defined as*

$$\begin{aligned} \underline{\mu}^{IB}(x) &= \inf_{\mathbf{a}^- \in \mathbb{R}^{p-1}} \inf_{\ell \in [-c,c]} E[g(x) | \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v], \\ \overline{\mu}^{IB}(x) &= \sup_{\mathbf{a}^- \in \mathbb{R}^{p-1}} \sup_{\ell \in [-c,c]} E[g(x) | \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_v] \end{aligned}$$

are found by solving two optimization problems

$$\begin{aligned} \underline{\mu}^{IB}(x) =& \hat{\mu}_0^{IB}(x) + \min_{\mathbf{a}^- \in \mathbb{R}^{p-1}} \left(\frac{\mathbf{a}^T A(x) \mathbf{a}}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}} - c \frac{|\mathbf{a}^T \hat{\mu}_1^{IB}(x)|}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}} \right), \\ \overline{\mu}^{IB}(x) =& \hat{\mu}_0^{IB}(x) + \max_{\mathbf{a}^- \in \mathbb{R}^{p-1}} \left(\frac{\mathbf{a}^T A(x) \mathbf{a}}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}} + c \frac{|\mathbf{a}^T \hat{\mu}_1^{IB}(x)|}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}} \right), \end{aligned}$$

where $\hat{\mu}_0^{IB}(x) = \mathbf{k}(x)^T K_v^{-1} \mathbf{y}$, $A(x) = (\mathbf{h}(x)^T - \mathbf{k}(x)^T K_v^{-1} H)^T \mathbf{y}^T K_v^{-1} H$ and $\hat{\mu}_1^{IB}(x) = \mathbf{h}(x)^T - \mathbf{k}(x)^T K_v^{-1} H$.

Proof. This result follows directly from proposition 5 by taking $p = 1$ and the single basis function $h_1(x) = h^{IB}(x, \mathbf{a}) = \mathbf{h}(x)\mathbf{a}$ so that

$$\begin{aligned} \hat{\mu}_0(x) =& \mathbf{k}(x)^T K_v^{-1} \mathbf{y} + \frac{\mathbf{a}^T (\mathbf{h}(x)^T - H^T K_v^{-1} \mathbf{k}(x))^T \mathbf{y} K_v^{-1} H \mathbf{a}}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}} \\ =& \hat{\mu}_0^{IB}(x) + \frac{\mathbf{a}^T A(x) \mathbf{a}}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}} \end{aligned}$$

and the bounds of $\hat{\mu}_1 \ell$ for $\ell \in [-c, c]$ are equal to

$$\pm c \frac{|\mathbf{a}^T (\mathbf{h}(x)^T - H^T K_v^{-1} \mathbf{k}(x))^T|}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}} = \pm c \frac{|\mathbf{a}^T \hat{\mu}_1^{IB}(x)|}{\mathbf{a}^T H^T K_v^{-1} H \mathbf{a}}.$$

□

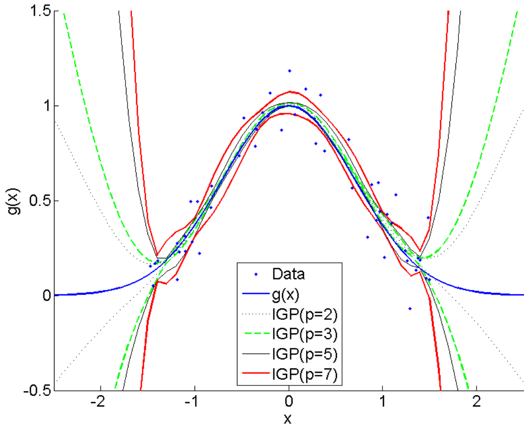

Figure 3: IGP-IB estimates of the function $g(x)$ given $n = 50$ observations using $c = 5$.

As the optimization problem in proposition 11 in general is not convex, computational costs can become significant as p increases.

Figure 3 shows, for different values of p, the posterior expectation for the IGP-IB model with $c = 5$ when and $\mathbf{h}(x) = [x^{p-1}, x^{p-2}, \ldots, x, 1]$. Notice that in this case, as the hyperparameters $\boldsymbol{\theta}_v$ are obtained as MAP estimate for the model with $h^{IB}(x, \mathbf{a}) = 1$, they depend only on the base kernel and thus are identical for all models, whatever is the number of basis functions p. Then, this model verifies the desirable property that its inferences always encompass those of the model with $p^* < p$, because all the priors considered in IGP-IB model with p^* basis functions are included also in the one with p basis function.

Since it does not assume a fixed set of prior basis function, this approach appears to be more appropriate in the lack of prior information. On the other side, it requires to solve a nonlinear optimization problem that may be questionable from the point of view of computational tractability. However, due to the strong influence on posterior inferences of the correlation model, further research should be devoted to this approach or similar ones that allow for partial (imprecise) elicitation of the basis functions or the GP kernel.

4. Application: hypothesis test for the equality of two functions

An equality test is used to detect differences between two regression functions $g_1(x)$ and $g_2(x)$ given the two independent samples $D_1 = (\mathbf{x}^{(1)}, \mathbf{y}^{(1)})$ and

$D_2 = (\mathbf{x}^{(2)}, \mathbf{y}^{(2)})$ of, respectively, n_1 and n_2 observations. Our aim is to extend the Bayesian test based on the GP presented in [13] using the IGP model. The approach in [13] assumes the same GP prior for the two functions g_1 and g_2 so that the two posterior distributions share the same hyperparameters. Here, we assume the same IGP set of priors $\mathcal{G}$ for the two functions, that is,

$$g_i(x) = f_i(x) + \mathbf{h}(x)\boldsymbol{w}^{(i)},$$

with $f_i(x) \sim GP(0, k_{\boldsymbol{\theta}})$, $p(\boldsymbol{w}^{(i)}) \propto \exp(\boldsymbol{\ell}^{(i)T}\boldsymbol{w}^{(i)})$ with $||\boldsymbol{\ell}^{(i)}||_\infty < c$, $i = 1, 2$. As a consequence, we are assuming that f_1 and f_2 are two GPs with the same kernel $k_{\boldsymbol{\theta}}(x, x')$ and the prior mean of $g_1(x)$ and $g_2(x)$ are obtained as a combination of the same set of basis functions. However, the coefficients $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ can be different. We assume that the two samples $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are affected by Gaussian noise with variance, respectively, σ_1^2 and σ_2^2. The hyperparameters $\boldsymbol{\theta}, \sigma_1, \sigma_2$ are obtained considering for both g_1 and g_2 the prior with $\ell^{(i)} = 0$. Then, after combining the two datasets $\{D_1, D_2\}$, we maximize the joint marginal probability of $(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \boldsymbol{\theta}, \sigma_1, \sigma_2)$ with respect to $\boldsymbol{\theta}, \sigma_1, \sigma_2$. Assuming that g_1 and g_2 are independent Gaussian processes, we have that

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \boldsymbol{\theta}, \sigma_1, \sigma_2) = p(\mathbf{y}^{(1)}|\mathbf{x}^{(1)}, \boldsymbol{\theta}, \sigma_1)p(\mathbf{y}^{(2)}|\mathbf{x}^{(2)}, \boldsymbol{\theta}, \sigma_2).$$

Then, the logarithm of the marginal likelihood of $\boldsymbol{\theta}, \sigma_1^2, \sigma_2^2$ is

$$\sum_{i=1}^{2} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \sigma_i)$$

where $\log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \sigma_i)$, is given in (14), up to an additive constant.

In the precise approach, given the prior $GP(0, k_{\boldsymbol{\theta}})$ for f_1 and f_2, we compute from (12) the posterior marginal distributions $p(\mathbf{g}_1^*|\mathbf{x}^*, D_1)$ and $p(\mathbf{g}_2^*|\mathbf{x}^*, D_2)$ of the functions $g_1(x)$ and $g_2(x)$ evaluated at the $m = n_1 + n_2$ test inputs $\mathbf{x}^* = \{x_1^{(1)}, \ldots, x_{n_1}^{(1)}, x_1^{(2)}, \ldots, x_{n_2}^{(2)}\}$. In this way, the equality of the two functions is tested at the covariates of the observations, that is, where we have the experimental evidence. Moreover, it is assumed that the observation covariates $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ cover the same region of the covariate space. This is done to avoid testing the equality in regions where there are no observations for one or both functions, as in these region we do not expect to be able to state any conclusion about equality or difference of the two functions. If applied in such regions, the precise test would always assign very large posterior probability to the hypothesis that there is no evidence of a difference between the functions. Using an IGP model, we can test the equality assumption in any subset $\mathcal{X}_T$ of the covariate space $\mathcal{X}$ by taking the m test inputs $\mathbf{x}^*$ so to cover uniformly the region of interest $\mathcal{X}_T$. If all priors in the IGP set entail the same decision, we retain it, if instead they lead to different decisions we conclude that a robust decision cannot be made in $\mathcal{X}_T$. This way, we can automatically identify a situation where data do not allow to state any conclusion.

Let us denote the mean and covariance functions of the posterior distributions of $g_1(x)$ and $g_2(x)$ as $\hat{\mu}^{(i)}(x)$ and $\hat{k}^{(i)}(x, x')$, $i = 1, 2$. Since the difference of two Gaussian variables is Gaussian, it follows that the posterior of the GP $\Delta g(x) = g_1(x) - g_2(x)$ is also a GP with mean and covariance functions $\Delta\hat{\mu}(x) = \hat{\mu}^{(1)}(x) - \hat{\mu}^{(2)}(x)$ and $\hat{k}_\Delta(x, x') = \hat{k}^{(1)}(x, x') + \hat{k}^{(2)}(x, x')$. Let $\Delta\mathbf{g}^*$, $\Delta\hat{\boldsymbol{\mu}}^*$ and $\hat{K}^*_\Delta$ be the difference $\Delta g(x)$, its mean and its covariance functions evaluated at the test covariates $\mathbf{x}^*$, then, we say that the two functions are equal with posterior probability $1-\alpha$ if the credible region for $\Delta\mathbf{g}^*$ includes the zero vector or, in other words, if:

$$(\Delta\hat{\boldsymbol{\mu}}^*)^T(\hat{K}^*_\Delta)^{-1}\Delta\hat{\boldsymbol{\mu}}^* \leq \chi^{-1}(1-\alpha|\nu), \tag{35}$$

where ν is the number of positive eigenvalues of $\hat{K}^*_\Delta$. In practice, as the number m of test inputs is likely to be considerably larger than the dimensionality of the GP representation, the matrix $\hat{K}^*_\Delta$ is not full rank. Thus, we decompose it as PDP^T, where D is the diagonal matrix of the eigenvalues $\lambda_1, \ldots, \lambda_m$ (sorted in descending order), and retain only the sub-matrices $P_\nu D_\nu P_\nu^T$ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_\nu$ which verify the condition $\lambda_{\nu+1}/\sum_{i=1}^m \lambda_i < \epsilon$, where ϵ is a small, positive constant. In the example below, we use $\epsilon = 0.0001$.

In the IGP model, the inference about $\chi^2_s(\boldsymbol{\ell}^{(1)}, \boldsymbol{\ell}^{(2)}) = (\Delta\hat{\boldsymbol{\mu}}^*)^T(\hat{K}^*_\Delta)^{-1}\Delta\hat{\boldsymbol{\mu}}^*$ depends on the choice of the prior, that is, on the values of $\boldsymbol{\ell}^{(1)}$, $\boldsymbol{\ell}^{(2)}$.

Proposition 12. *The IGP model is a prior ignorance model for inferences about* χ^2_s*, i.e.,*

$$\underline{\chi}^2_s = 0 \quad \overline{\chi}^2_s = +\infty.$$

Proof. A priori

$$\chi^2_s = (E[\boldsymbol{w}^{(1)}] - E[\boldsymbol{w}^{(2)}])^T H^T (K^{(1)**} + K^{(2)**})^{-1} H (E[\boldsymbol{w}^{(1)}] - E[\boldsymbol{w}^{(2)}])$$

Then, the lower bound is found by assuming the same prior for $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$, and the upper assuming the priors $p(\boldsymbol{w}^{(i)}) \propto \exp(\boldsymbol{\ell}^{(i)T}\boldsymbol{w}^{(i)}),\ i = 1, 2$, with, for instance, $\ell_j^{(1)} > 0$ and $\ell_j^{(2)} < 0$ for all $j = 1, \ldots, p$. □

A posteriori, let

$$\Delta\hat{\boldsymbol{\mu}}_0^* = \hat{\boldsymbol{\mu}}_0^{*(1)} - \hat{\boldsymbol{\mu}}_0^{*(2)}$$

$$\Delta\hat{\boldsymbol{\mu}}_1^*(\boldsymbol{\ell}^{(1)}, \boldsymbol{\ell}^{(2)}) = \hat{\boldsymbol{\mu}}_1^{*(1)}\boldsymbol{\ell}^{(1)} - \hat{\boldsymbol{\mu}}_1^{*(2)}(\mathbf{x}^*)\boldsymbol{\ell}^{(2)}$$

The lower (upper) bounds for $\chi_s^2(\boldsymbol{\ell}^{(1)}, \boldsymbol{\ell}^{(2)})$ are obtained by minimizing (maximizing) w.r.t. $\boldsymbol{\ell}^{(1)}, \boldsymbol{\ell}^{(2)} \in \mathbb{L}$ the statistic:

$$\chi_s^2 = (\Delta\hat{\boldsymbol{\mu}}_0^* + \Delta\hat{\boldsymbol{\mu}}_1^*(\boldsymbol{\ell}^{(1)}, \boldsymbol{\ell}^{(2)}))^T(\hat{K}_\Delta^*)^{-1}(\Delta\hat{\boldsymbol{\mu}}_0^* + \Delta\hat{\boldsymbol{\mu}}_1^*(\boldsymbol{\ell}^{(1)}, \boldsymbol{\ell}^{(2)})). \qquad (36)$$

Notice that this is a quadratic programming problem, as χ_s^2 can be rewritten as

$$\begin{aligned}\chi_s^2 =&(\Delta\hat{\boldsymbol{\mu}}_0^* + \hat{\boldsymbol{\mu}}_1^{*(1,2)}\boldsymbol{\ell}^{(1,2)})^T(\hat{K}_\Delta^*)^{-1}(\Delta\hat{\boldsymbol{\mu}}_0^* + \hat{\boldsymbol{\mu}}_1^{*(1,2)}\boldsymbol{\ell}^{(1,2)})\\ =&\Delta\hat{\boldsymbol{\mu}}_0^{*T}(\hat{K}_\Delta^*)^{-1}\Delta\hat{\boldsymbol{\mu}}_0^* + \boldsymbol{\ell}^{(1,2)T}Q\boldsymbol{\ell}^{(1,2)} + \mathbf{c}^T\boldsymbol{\ell}^{(1,2)}.\end{aligned}$$

where $Q = \hat{\boldsymbol{\mu}}_1^{*(1,2)T}(\hat{K}_\Delta^*)^{-1}\hat{\boldsymbol{\mu}}_1^{*(1,2)}$, $\mathbf{c}^T = 2\Delta\hat{\boldsymbol{\mu}}_0^{*T}(\hat{K}_\Delta^*)^{-1}\hat{\boldsymbol{\mu}}_1^{*(1,2)}$, $\boldsymbol{\ell}^{(1,2)} = [\boldsymbol{\ell}^{(1)T}, \boldsymbol{\ell}^{(2)T}]^T$ and $\hat{\boldsymbol{\mu}}_1^{*(1,2)} = [\hat{\boldsymbol{\mu}}_1^{*(1)},\ -\hat{\boldsymbol{\mu}}_1^{*(2)}]$.

4.1. Numerical example

Let us consider two samples D_1 and D_2 that we wish to compare on the subset $\mathcal{X}_T = [a, b]$ of the covariate space. Assuming an observation noise $v \sim \mathcal{N}(0, \sigma_v = 0.2)$, we sample D_1 and D_2 from:

Case A: $x_i^{(1,2)} \sim U[-2, 2],\ \ y_i^{(1,2)} = f(x_i) + v_i,$

Case B: $x_i^{(1)} \sim U[-2, 2],\ \ y_i^{(1)} = f(x_i) + v_i,$
$x_i^{(2)} \sim U[-2, 2]\ \ \ y_i^{(2)} = g(x_i) + v_i,$

Case C: $x_i^{(1)} \sim U[-2, 0],\ \ y_i^{(1)} = f(x_i) + v_i,$
$x_i^{(2)} \sim U[-2, 4],\ \ y_i^{(2)} = g(x_i) + v_i,$

Case D: $x_i^{(1)} \sim U[-2, 2],\ \ y_i^{(1)} = f(x_i) + v_i,$
$x_i^{(2)} \sim U[-2, 4],\ \ y_i^{(2)} = g(x_i) + v_i,$

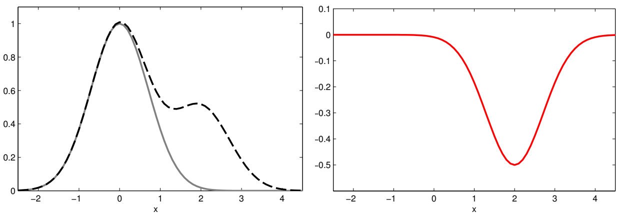

Figure 4: Left: Functions f (continuous line) and g (dashed line). Right: difference $f - g$.

where $f(x) = \exp(-x^2)$ and $g(x) = f(x) + 0.5f(x-2)$ (see Figure 4). For each scenario the two datasets D_1 and D_2 have been simulated only once. We have tested the difference between the two samples for different test subsets $\mathcal{X}_T \in [-2, b]$. The difference $f(x) - g(x)$ is about zero for $x < 0$, is large ($> \sigma_v$) in the interval $[1, 3]$ and is small ($< \sigma_v$) in $[0, 1]$. Therefore, we expect to easily detect a difference between the two samples when $b > 1$, whereas for $b < 1$ the decision is more difficult and for $b < 0$ we can assume that the two functions are equal. Table 1 shows the decisions for the precise and the imprecise tests with constant mean function at different values of c and b. One can notice that for $c = 10$ we are most often undecided (save when the decision is simple, e.g., in cases B and D when $b > 1$ and thus all tests recognize the difference) as the imprecision is very large in this case.

On the other side, for $c = 1$ the test makes almost always the same decision as the precise test, as the imprecision is very small in this case. When $c = 5$ we have a better balance between robustness and power: the IGP test makes the same decision as the precise one when there is enough information to make a robust decision, whereas it is undecided when the decision is difficult due to the lack of information. For instance, in case A with $b = 2$ the precise test always issues a no difference decision. The same happens in case C, although the two situations are very different, because in the first case $f_1 = f_2$ and we can observe both functions on the entire set $\mathcal{X}_T$, whereas in the second case $f_1 \neq f_2$ but we cannot see it as we observe f_1 only in the range $[-2, 0]$ where the two function are almost identical. On the other side, the imprecise test detects the difference of the two situations, and in case A it correctly issues a no difference decision, whereas in case C it is undecided, thus acknowledging that there is not enough information to make a decision. Something similar can be observed also in case D: when $b = 0$ both the precise and imprecise tests issue a no difference decision as in this range the two function can be actually considered identical; when, instead,

		GP		IGP	
Case	b	n=50	n=200	n=50	n=200
A	2	0	0	0/0/2	0/0/2
A	4	0	0	0/2/2	0/2/2
B	0	0	0	0/0/2	0/0/2
B	1	0	1	0/2/2	1/1/1
B	2	1	1	1/1/1	1/1/1
B	4	1	1	1/1/1	1/1/1
C	0	0	0	0/0/2	0/0/2
C	1	0	0	0/0/2	0/0/2
C	2	0	0	0/2/2	0/2/2
C	4	0	0	0/2/2	0/2/2
D	0	0	0	0/0/0	0/0/2
D	1	0	1	2/2/2	1/1/1
D	2	1	1	1/1/1	1/1/1
D	4	1	1	1/1/1	1/1/1

Table 1: Decisions of the precise test for $c = 1/5/10$, where 0 indicates that the two functions are equal with posterior probability $1 - \alpha$, 1 indicates that the two functions are different (i.e., the posterior probability that they are equal is less than α), 2 indicates indecision (i.e., the decision depends on the prior).

$b = 1$, the functions are different, but, since the difference is small, it cannot be clearly detected with only $n = 50$ data. However, the imprecise test recognizes that the decision is somehow difficult and is undecided, whereas the precise test can only decide that there is no difference. For $n = 200$, the information is enough to make both tests detect a difference.

5. Conclusions

In this paper we have presented a model of prior near ignorance about the value of a regression function based on the Gaussian process. With respect to the set of minimal properties that the lower and upper expectation of a prior-near ignorance model should verify according to [7] we can state what follows.

- Translation invariance is verified by the IGP set of priors, and this implies that posterior imprecision does not depend on translation of the sample mean.
- The IGP with p basis functions models prior ignorance about the expectation of $g(x)$ conditional on its values at a maximum of $p-1$ covariates.
- The IGP model verifies learning. Moreover, posterior inferences are tractable as a closed-form expression has been provided for the upper and lower bounds of $E[g(x)]$.
- Convergence has been proven in the case of linear regression in the set of basis functions $\mathbf{h}(x)$.

We have shown that the IGP model can be used to make inferences about the regression functions which are more robust with respect to the choice of the prior. In fact, for those subsets of $\mathcal{X}$ where there are many observations inferences are very close with those of the precise model, whereas in those subsets with no observations the imprecision of the prediction can be very high, thus reflecting the actual lack of knowledge. As a consequence of this, decisions based on this model should be considered more reliable. For instance, we have applied the IGP to test the difference between regression functions, and shown that the IGP model allows us to acknowledge when the available data are not informative enough to make a robust decision.

Although in this paper we have only considered univariate functions, the c-IGP model can be extended to the multivariate case where x is a vector of covariates.

Moreover, as a strong prior information is introduced in the IGP regression by the choice of a model for the covariance between functions values at any two covariates (that is, by the choice of the base kernel and the set of basis functions), we have proposed to use a set of IGP models with (single) basis function free to vary in a set of functions. This and similar approaches allowing for a weaker specification of the covariance function should be considered in future research.

In this work we have focused on GP-based methods to develop an approach to prior near-ignorance modeling in nonparametric regression. This approach has the potential to be extended, in future work, to other regression techniques, taking advantage also of the connections they have with GPs [4, Sec. 6].

Acknowledgements

The author would like to thank Alessio Benavoli for his valuable comments and suggestions.

References

[1] A. O'Hagan, Curve fitting and optimal design for prediction, Journal of the Royal Statistical Society. Series B (Methodological) 40 (1) (1978) 1–42.

[2] R. M. Neal, Regression and classification using gaussian process priors, in: Bernardo, et al. eds., Bayesian Statistics 6: Proc of the 6th Valencia international meeting, Vol. 6, 1998, p. 475.

[3] D. J. MacKay, Introduction to Gaussian processes, In Bishop, C. M., editor, Neural Networks and Machine Learning (1998) 133–166.

[4] C. E. Rasmussen, C. Williams, Gaussian processes for machine learning, The MIT Press, Cambridge, MA, USA, 2006.

[5] C. E. Rasmussen, The Gaussian Processes Web Site, http://www.gaussianprocess.org/ (February 2011).

[6] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, Bayesian data analysis, CRC press, 2013.

[7] A. Benavoli, M. Zaffalon, Prior near ignorance for inferences in the k-parameter exponential family, Statistics 49 (5) (2015) 1104–1140. doi:10.1080/02331888.2014.960869.
URL `http://www.idsia.ch/ alessio/benavoli2014b.pdf`

[8] J. O. Berger, An overview of robust bayesian analysis with discussion, Test 3 (1) (1994) 5–124.

[9] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, New York, 1991.

[10] T. Augustin, F. P. Coolen, G. de Cooman, M. C. Troffaes, Introduction to imprecise probabilities, John Wiley & Sons, 2014.

[11] L. R. Pericchi, P. Walley, Robust Bayesian credible intervals and prior ignorance, International Statistical Review 58 (1991) 1–23.

[12] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J of the Royal Statistical Society. Series B (Methodological) 58 (1) (1996) 3–57.

[13] A. Benavoli, F. Mangili, Gaussian Processes for Bayesian hypothesis tests, in: Proc 18th AIstat Conference, Society for Artificial Intelligence and Statistics, 2015.

[14] F. Mangili, A prior near-ignorance gaussian process model for nonparametric regression, in: T. Augustin, S. Doria, E. Miranda, E. Quaeghebeur (Eds.), ISIPTA '15: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications, pp. 187–196.
URL `http://www.sipta.org/isipta15/data/paper/15.pdf`